

CONTROVERSIA SOBRE MENTES Y MÁQUINAS

Edición de Alan Ross Anderson



Adviértase lo difícil que resultaría admitir, tras observar cierta hipotética máquina, que ésta «ha elegido» o se «ha enamorado» Las mismas pruebas que aceptaríamos como flagrantes en un ser humano, serían, en el caso de las máquinas, objeto de controversia sin fin. Algunas ventajas tenemos, sin embargo, sobre los pensadores precibernéticos. ventajas que residen en el estado actual de la teoría y la práctica de la cibernética y la matemática, de sus estímulos y de sus miedos A partir de ciertas definiciones nuevas, insoslayables, puede argumentarse sobre la posibilidad de que una máquina llegue a traspasar una. predefinida, frontera del pensamiento.

Ensayos sobre la inteligencia artificial, como el de J. R. Lucas, «Mentes,,máquinas y Gödel», y el de A. M. Turing, «Maquinaria computadora e inteligencia», recogidos en este libro, son ya piezas clásicas. Lucas se empeña en demostrar que una máquina jamás puede pensar, que no puede reducirse un cerebro a una estructura mecánica. Turing, por el contrario, intenta probar la hipótesis opuesta, desarmando todas las objeciones que acostumbran esgrimirse contra la capacidad de pensar de una máquina: objeciones teológicas, las llamadas del «avestruz» («¿máquinas que piensan?, ¡horror!»), las matemáticas, las objeciones sobre incapacidades varias (la de que una máquina sea chistosa, cariñosa, altruista, etc.), las objeciones sobre la analogía con el sistema nervioso humano, las de la percepción extrasensorial, y muchas más. En torno a estas dos posiciones se organizan las demás, matizadas aquí por otros científicos y pensadores como Hilary Putman y. sobre todo, Paul Ziff, del que reproducimos el brillante ensayo «El sentir de los robots», donde, a la pregunta convencional «¿Siente un robot?», afirma que no, porque, entre otras razones, «... desear el robot del prójimo sería como desear su coche y no a su mujer».

No en vano este libro, Controversia sobre mentes y máquinas, fue concebido y editado por un doctor en Filosofía, profesor en la Universidad de Yale, Estados Unidos. Alan Ross Anderson, coeditor también, con Maurice Mandelbaum y Francis Gramlich. del libro «Philosophic Problems», Decimos bien no en vano porque, una vez más, cuestiones que aparentemente deberían ser competencia tan sólo de la ciencia están suscitando interrogantes metafísicos, religiosos morales, etc., que. de hecho, suscitan el interés no sólo ya de los especialistas, sino de todos aquellos que se plantean va la rebelión de las máquinas, ya sea hacia la chatarra. o hacia la genialidad.



Alan Ross Anderson & A. M. Turing & Michael Scriven & J. R.
Lucas & Keith Gunderson & Hilary Putnam & Paul Ziff & J. J. C.
Smart & Ninian Smart

Controversias sobre mentes y máquinas

Metatemas - 7

ePub r1.0
oronet 20.12.2020

Título original: *Mind and machines*

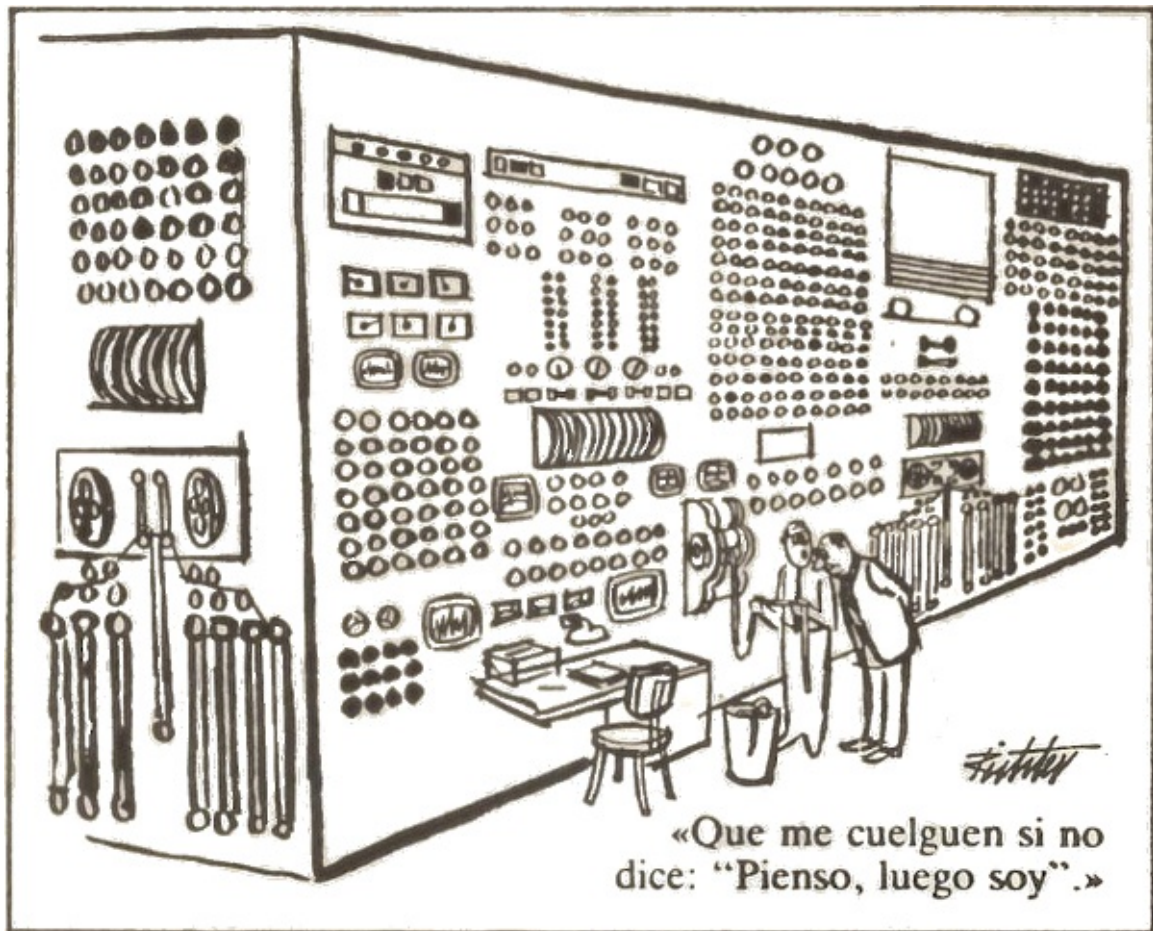
Alan Ross Anderson & A. M. Turing & Michael Scriven & J. R. Lucas & Keith Gunderson & Hilary Putnam & Paul Ziff & J. J. C. Smart & Ninian Smart, 1964

Traducción: Francisco Martín

Editor digital: oronet

ePub base r2.1





«Que me cuelguen si no
dice: "Pienso, luego soy".»

Introducción

Alan Ross Anderson

Todo es incierto con respecto al desarrollo definitivo de la conciencia mecánica, a juzgar por las máquinas actuales que poseen poca conciencia... Incluso una patata en una cueva oscura posee cierta astucia que le es de suma utilidad

Samuel Butler, *Erewhon*

En los últimos años el desarrollo de computadoras electrónicas ha dado un nuevo giro a la problemática de las relaciones entre acontecimientos «mentales» y «mecánicos», suscitando una interminable controversia. A partir de 1950 se han publicado más de 1.000 trabajos sobre la cuestión de si las «máquinas» pueden o no pueden «pensar».

Ante tan abrumador volumen de trabajos, cualquier selección de artículos de actualidad, dedicada a constituir una pequeña antología como la presente, refleja necesariamente, y hasta cierto punto, las preferencias personales del editor (aunque, en mi caso, no pida excusas por incluir el artículo del fallecido A. M. Turing, que en un plazo de trece años se ha convertido en un clásico en la materia). Sí, en cambio, me excuso ante aquellos amigos, cuyos consejos he desoído, y ante aquellos otros que me confiaron artículos interesantes, cuya inclusión ha sido imposible. Trataré a continuación de exponer los principios por los que me he guiado para la selección de los artículos, aunque quizá se entienda mejor si hago un repaso de los problemas filosóficos en cuestión.

¿Cómo reconocer una *entidad* (por utilizar un término que no prejuzgue el aspecto de si el sujeto es un organismo o una máquina) dotada de *mente*? Todos sabemos que las *personas* tienen mente y pueden pensar; también los monos piensan, en el sentido de que resuelven cierta clase de problemas sencillos, del mismo modo que las ratas. Moscas y mosquitos también parecen obtener una calificación aceptable, pero la mayoría de nosotros no consideramos que tengan mente (en sentido estricto). Independientemente de lo que entendamos por «tener mente», la mayoría de la gente coincidirá seguramente en que una caja registradora carece de ella, y también la más perfecta de las computadoras actuales. Pero ¿dónde está exactamente la

frontera? Las dos posiciones que exponemos a continuación representan los dos extremos entre los que se entabla la discusión más al día sobre el tema:

1. Podría decirse que los seres humanos no son más que elaboradísimos mecanismos de relojería y que nuestra facultad de tener «mente» es sencillamente consecuencia de que esta relojería sea *muy* elaborada, o bien,

2. podría decirse que *cualquier* máquina es un simple producto del ingenio humano (no muy distinta en principio a una pala) y que, aunque nosotros tengamos *mente*, no podemos impartir esa característica peculiar a ninguna otra cosa, salvo a nuestra prole. Ninguna máquina puede asumir una característica tan exclusivamente humana.

Naturalmente, todos «sabemos» que los seres humanos poseen mente y que las computadoras (por ahora) no; el problema filosófico ha venido a centrarse en parte en torno a lo que queremos *significar* cuando argumentamos de este modo. Dicho de otra manera: ¿qué tipo de evidencia nos impulsa a decir que una máquina tiene mente, piensa o puede sentir? Admitir que una máquina posee semejantes atributos conlleva probablemente el compromiso de referirse a la máquina del mismo modo en que nos referimos al hablar de entidades que irrefutablemente poseen estos atributos, es decir, nosotros mismos. [«No le (¿lo?) molestes; está pensando». «Está (¿la máquina?) de mal humor esta mañana». Y así sucesivamente].

Visto así, hablar de este modo de las máquinas parece absurdo, pero existen otros modos de considerar el problema. Está claro que las máquinas no pueden, por ahora, actuar en ciertos aspectos como seres humanos, pese a que sean capaces de realizar muchas cosas tan bien o mejor que las personas. Las computadoras pueden jugar al *nim*, o al *pennymatching*, tan bien como nosotros, pueden calcular con mayor rapidez y vuelan mejor. ¿Qué razones hay para pensar que en principio es imposible construir máquinas que sean también capaces de pensar, sentir, tener dudas, etc.? Si negáramos esta posibilidad, ¿cómo podríamos tener la certeza de que nuestros argumentos son racionales (y no exponentes de un increíble espíritu de intolerancia u hostilidad por nuestra parte)?

Valga lo dicho cuando menos como ejemplo del tipo de problemas que se discuten en los artículos reunidos en este libro. Ninguno de ellos trata exclusivamente de la filosofía de la mente, en el sentido tradicional, y ninguno gira en torno a problemas técnicos de ingeniería electrónica: nuestro propósito ha sido más bien insistir sobre la conjunción «y» de *Mentes y máquinas*. Por ello he tenido que excluir necesariamente muchos trabajos excelentes que, aunque vinculados al tema, estaban escritos desde una perspectiva filosófica con interés marginal para las máquinas, o desde el punto de vista del matemático o del ingeniero, pero con interés marginal para la filosofía de la mente. Se incluye al final una bibliografía de utilidad para el lector que desee ampliar el tema en ambas direcciones.

El segundo principio selectivo responde, a mi entender, a que la filosofía es algo fundamentalmente dialéctico. Así pues, se han elegido los artículos con el fin de

ilustrar cierta interacción de los autores, al objeto de componer un diálogo. Aunque a uno no le interese intervenir en la polémica, se puede aprender bastante sobre lo que se dirime contemplando (por así decirlo) a los diversos contendientes en liza.

Quedo agradecido a los autores y editores de las publicaciones en que aparecieron originalmente los ensayos por su amabilidad al autorizarme su reedición. Quedo igualmente agradecido a Keith Gunderson y a Wesley Salmón por sus valiosas indicaciones, y a Marvin Minsky por cederme un ejemplar de su trabajo *Steps Toward Artificial Intelligence* (ver bibliografía) que me fue de gran utilidad, tanto desde la perspectiva conceptual como bibliográfica. Doy finalmente las gracias a mi esposa, no sólo por aportar su Mente, sino también por ayudarme con una Máquina en la preparación del manuscrito.

Maquinaria computadora e inteligencia^[*]

A. M. Turing

1. El juego de imitación

Propongo que consideremos la siguiente pregunta: «¿Pueden pensar las máquinas?» para empezar, definamos el significado de los términos «máquina» y «pensar», pero es una actitud peligrosa. Si hemos de llegar al significado de las palabras «máquina» y «pensar» a través de su utilización corriente, difícilmente escaparíamos a la conclusión de que hay que buscar el significado y la respuesta de la pregunta «¿Pueden pensar las máquinas?» mediante una encuesta tipo Gallup. Pero es absurdo. En lugar de intentar tal definición, sustituiremos la pregunta por otra estrechamente relacionada con ella y que se expresa con palabras relativamente inequívocas.

El problema en su nuevo planteamiento puede exponerse en términos de un juego que denominaremos «juego de imitación». Intervienen en él tres personas: un hombre (A), una mujer (B) y un preguntador (C), indistintamente de uno u otro sexo. El preguntador se sitúa en una habitación aparte y, para él, el juego consiste en determinar quién de los otros dos es el hombre y quién la mujer. Los conoce por la referencia X e Y, y al final del juego determina si «X es A e Y es B» o si «X es B e Y es A». El preguntador puede plantear a A y a B preguntas como éstas: «Por favor X, ¿podría decirme cuán largo es su pelo?»

Supongamos que X es realmente A, entonces es A quien contesta. El objetivo de A en el juego es lograr que C efectúe una identificación errónea, por lo que su respuesta podría ser: «Mi pelo es corto, escalonado, y los mechones más largos son de unos veinte centímetros».

Para que el preguntador no se guíe por el timbre de voz, las respuestas deben ir por escrito o, mejor aún, mecanografiadas. Lo ideal es disponer de un impresor telegráfico que comunique las dos habitaciones.

Otro procedimiento consiste en que un intermediario repita pregunta y respuesta. El objeto del juego para el tercer jugador (B) es ayudar al preguntador. La mejor estrategia para la jugadora es probablemente responder la verdad, añadiendo quizás a sus respuestas cosas como ésta: «¡Soy la mujer, no le haga caso!», pero de nada sirve, ya que el hombre puede hacer observaciones similares.

Ahora planteemos la pregunta: «¿Qué sucede cuando una máquina sustituye a A en el juego?» ¿Se pronunciará el preguntador en este caso tan erróneamente como lo hace cuando en el juego participan un hombre y una mujer? Estas preguntas sustituyen a la original: «¿Pueden pensar las máquinas?»

2. Crítica del nuevo problema

Del mismo modo que preguntamos: «¿Cuál es la respuesta a este nuevo tipo de pregunta?», podemos preguntar: «¿Merece la pena resolver esta nueva pregunta?» Resolvamos esta última pregunta sin plantear más objeciones para cortar una regresión infinita.

El nuevo problema presenta la ventaja de que traza una línea definida entre las aptitudes físicas e intelectuales de una persona. Ningún ingeniero o químico puede atribuirse la capacidad de producir un material que no pueda distinguirse de la piel humana. Quizá sea posible algún día, pero, aun suponiendo la viabilidad de semejante invención, nos parece que de poco serviría tratar de hacer una «máquina pensante» más humana, forrándola con esa epidermis artificial. El modo en que hemos planteado el problema refleja el obstáculo que impide al preguntador ver o tocar a los otros concursantes, oír su voz. Otras ventajas del criterio propuesto pueden resumirse en un modelo de preguntas y respuestas. Por ejemplo:

P: Por favor, escriba un soneto sobre el tema del Cuarto Puente.

R: Hágame otra pregunta; la poesía no es mi fuerte.

P: Sume 34.957 con 70.764.

R: (Pausa de unos 30 segundos) 105.621.

P: ¿Juega al ajedrez?

R: Sí.

P: Tengo el rey en la casilla IR y ninguna otra pieza. Usted tiene sólo el Rey en la casilla 6R la Dama en 1D. Le toca mover. ¿Qué juega?

R: (Pausa de unos 15 segundos) La Dama a D8, mate.

El método de preguntas y respuestas parece adecuado para introducir casi todos los campos de actividad humana que queramos. No vamos a sancionar a la máquina por su incapacidad para destacar en concursos de belleza, del mismo modo que no castigamos a una persona por perder una carrera en una competición aérea. Las condiciones del juego hacen irrelevantes esas torpezas. Los «testigos» pueden alardear, si lo creen conveniente, tanto como deseen con respecto a sus encantos, su fuerza o su heroísmo, pero el preguntador no puede exigir demostraciones fehacientes.

El juego quizá provoque críticas porque la máquina tiene demasiados factores en contra. Si una persona lo intentara haciéndose pasar por la máquina, sin duda haría un papel deplorable. Quedaría rápidamente eliminada por lentitud e inexactitud aritmética. ¿No harán las máquinas algo que permita la definición de pensamiento,

pero que es muy distinto a lo que hace una persona? Se trata de una objeción de peso, pero cuando menos podemos decir que, dado que es posible construir una máquina que realice satisfactoriamente el juego de imitación, la objeción no viene al caso.

Podría alegarse que la mejor estrategia en el «juego de imitación», para la máquina, es posiblemente algo distinto a la imitación de la conducta humana. Puede, pero yo no creo que esto influya demasiado. En cualquier caso, no nos proponemos aquí analizar la teoría del juego y supondremos que la mejor estrategia es tratar de dar las respuestas que una persona daría con toda naturalidad.

3. Las máquinas que intervienen en el juego

La cuestión que planteábamos en el apartado 1 carece de precisión si no especificamos qué entendemos por el término «máquina». Es lógico que deseemos que nuestras máquinas estén dotadas de cualquier tipo de ingeniería mecánica. Del mismo modo que aceptamos la posibilidad de que un ingeniero o un equipo de ingenieros construya una máquina que funcione, pero cuya modalidad operacional no pueden describir satisfactoriamente sus constructores porque se han servido de un método fundamentalmente experimental. Finalmente, excluirémos de la categoría de máquinas a las personas nacidas del modo habitual. Es difícil adaptar las definiciones de modo que cumplan estos tres requisitos. Se puede insistir, por ejemplo, en que el equipo de ingenieros sea de un solo sexo, lo cual no sería satisfactorio, ya que probablemente se puede crear un individuo completo a partir de una simple célula epidérmica de un hombre (pongamos por caso). Esto sería una proeza de biogenética merecedora de máxima admiración, pero no por ello la calificaríamos de «construcción de máquina pensante». Esto nos obliga a descartar el requisito de permitir cualquier tipo de técnica, y con mayor razón dado que el interés actual por las «máquinas pensantes» se ha suscitado gracias a un tipo particular de máquina, generalmente denominada «computadora electrónica» o «computadora digital». Con arreglo a esto, sólo permitiremos que tomen parte en el juego las computadoras digitales.

A primera vista esta limitación parece muy drástica, pero intentaré demostrar que no es así. Para ello es necesario un breve resumen sobre la naturaleza y las propiedades de estas computadoras.

Podría también aducirse que esta identificación de las máquinas con las computadoras digitales, al igual que nuestro criterio sobre el término «pensar», son insatisfactorias si (en contra de lo que creo) resulta que las computadoras digitales son incapaces de hacer un buen papel en el juego.

Existen ya varias computadoras operacionales, y es lógico que se diga: «¿Por qué no realizar el experimento ahora mismo? No resultaría difícil cumplir los requisitos del juego. Se pueden utilizar varios preguntadores, compilando unas estadísticas para comprobar cuántas veces se produce la identificación correcta». La respuesta

inmediata es que no se trata de plantearse si todas las computadoras digitales actuarán bien en el juego, ni de si las actuales computadoras actuarán bien, sino de si existen computadoras imaginables que actúen bien. Pero esto es sólo la respuesta inmediata, más adelante consideraremos la cuestión bajo otra perspectiva.

4. Computadoras digitales

Podemos explicar el concepto de computadoras digitales diciendo que son unas máquinas ideadas para realizar cualquier tipo de operación propia de un computador humano. El computador humano sigue unas reglas determinadas sin opción a desviarse de ellas bajo ningún concepto. Supongamos que esas reglas figuran en un libro que cambia cada vez que el computador acomete un nuevo trabajo. Dispone también de una cantidad ilimitada de papel para efectuar cálculos y hace las multiplicaciones y sumas pertinentes con una «máquina de bolsillo», pero esto no tiene importancia.

Si utilizamos como definición la anterior explicación, corremos el riesgo de caer en una argumentación circular. Para evitarlo, esbozaremos los medios con los que se logra el efecto deseado. Suele considerarse que una computadora digital consta de tres partes:

- a. almacenamiento
- b. unidad procesadora
- c. control

El almacenamiento es el acopio de información y corresponde al papel sobre el que se efectúa la computación humana, ya sea el papel en que la persona realiza los cálculos o aquél en el cual está impreso el libro de reglas. Del mismo modo que el computador humano efectúa sus cálculos con su cabeza, parte del almacenamiento corresponde a la memoria de la máquina.

La unidad procesadora es el sector que realiza las distintas operaciones de cálculo. La naturaleza de estas operaciones varía de una máquina a otra. Generalmente pueden efectuar operaciones bastante largas, tales como «Multiplicar 3540675445 por 7076345687», pero en algunas máquinas sólo pueden llevarse a cabo operaciones muy simples, tales como «Escribe 0».

Hemos mencionado que el «libro de reglas», de que se vale el computador, se sustituye en la máquina por una parte del almacenamiento. Esta se denomina «tabla de instrucciones». Corresponde al control comprobar que las instrucciones se sigan correctamente y en su debido orden. El control está construido de tal manera que es infalible.

La información almacenada suele estar dividida en paquetes de tamaño relativamente modesto. En una máquina concreta, por ejemplo, el paquete puede

constar de diez dígitos decimales. Se asignan números a las partes del almacenamiento en que se guardan los diversos paquetes de información, con arreglo a una modalidad sistemática. Un ejemplo de instrucción corriente podría ser: «Suma la cifra almacenada en la posición 6809 a la situada en la 4302 y devuelve el resultado de la última posición de almacenamiento». Ni que decir tiene que la operación no se desarrolla en la máquina expresada de este modo, sino que se lleva a cabo siguiendo una codificación como 6809430217. La cifra 17 indica cuál de las posibles operaciones hay que efectuar con las dos cifras. En cuyo caso la operación es la anteriormente descrita: «Suma la cifra...». Se advertirá que la instrucción consta de diez dígitos y, por lo tanto, constituye exactamente un paquete informativo. El control suele captar las instrucciones a seguir en el orden de posición en que están almacenadas, aunque a veces pueda surgir una instrucción como ésta: «Sigue ahora la instrucción almacenada en la posición 5606 y continúa», o bien: «Si la posición 4505 contiene 0, sigue la instrucción almacenada en 6707; en caso contrario continúa».

Las instrucciones de este tipo son muy importantes porque permiten la repetición de una secuencia de operaciones una y otra vez hasta que se cumple un determinado requisito, pero, al hacerlo, la máquina sigue en cada repetición, no nuevas instrucciones, sino las mismas indefinidamente. Recurramos a una analogía casera: supongamos que mamá desea que Tommy pase por el zapatero cada mañana camino del colegio para ver si han arreglado sus zapatos; puede decírselo cada mañana, o puede dejar una nota permanente en el vestíbulo para que el niño la vea al salir y recuerde que tiene que pasar por el zapatero, y luego, al volver, si trae los zapatos, rompa la nota. El lector debe aceptar como un hecho la construcción de computadoras digitales que, efectivamente, se han construido con arreglo a los principios expuestos y que realmente mimetizan con gran fidelidad los actos de un computador humano.

El libro de reglas que, según hemos señalado, utiliza el computador humano es, naturalmente, una ficción convencional. Los computadores humanos recuerdan en realidad lo que tienen que hacer. Si queremos hacer una máquina que mimetice el comportamiento de un computador humano en operaciones complicadas, hay que preguntarle a éste cómo lo hace y luego transferir la respuesta en forma de tabla de instrucciones. La elaboración de tablas de instrucciones suele denominarse «programación». La «programación de una máquina para que efectúe la operación A» significa insertar en la máquina la tabla de instrucción adecuada para que lleve a cabo A.

Una variante interesante de la idea de computadora digital es la «computadora digital con un elemento aleatorio». Estas máquinas disponen de instrucciones en las que interviene un dado o un proceso electrónico equivalente; una instrucción de este tipo puede ser, por ejemplo: «Arroja el dado y almacena la cifra resultante en 1000». A veces se las denomina máquinas de libre voluntad (aunque personalmente yo no utilice esta expresión). Normalmente no se puede determinar por simple observación

de la máquina si ésta posee un elemento aleatorio, ya que se logra un efecto similar con dispositivos cuya elección depende de los dígitos de los decimales de π .

La mayoría de las computadoras digitales poseen un almacenamiento finito, aunque no existe dificultad teórica en la concepción de una computadora de almacenamiento ilimitado. Naturalmente, sólo podría utilizarse una parte finita de cada fase. De igual modo se habría podido construir una cantidad finita, pero cabe imaginar que sucesivamente fueran añadiéndose otras. Estas computadoras presentan especial interés teórico y las denominaremos computadoras de capacidad infinita.

El concepto de computadora digital es antiguo. Charles Babbage, profesor de matemáticas en la Universidad de Cambridge entre 1828 y 1839 concibió una a la que denominó Máquina Analítica, pero no la terminó. Aunque Babbage expuso los principios fundamentales, la máquina no representaba en aquella época gran interés. Su rapidez habría sido mucho mayor que la de un computador humano, pero unas 100 veces inferior a la de la máquina de Manchester^[2], que a su vez es una de las máquinas modernas más lentas. El almacenamiento era puramente mecánico y se efectuaba por medio de ruedas y tarjetas.

El hecho de que la Máquina Analítica de Babbage estuviera concebida de forma totalmente mecánica nos ayudará a despejar cualquier superstición. Muchas veces se atribuye importancia al hecho de que las computadoras digitales modernas son eléctricas, igual que el sistema nervioso. Como la máquina de Babbage no era eléctrica, y como todas las computadoras digitales son en cierto modo equivalentes a ella, el empleo de la electricidad no es teóricamente relevante.

Siempre que se trata de señalización rápida interviene, claro, la electricidad. Por lo tanto, no es de extrañar que ésta se halle relacionada con ambos conceptos. En el sistema nervioso los fenómenos químicos son, cuando menos, tan importantes como los eléctricos. En ciertas computadoras el sistema de almacenamiento es fundamentalmente acústico. Por lo tanto, el empleo de la electricidad como propiedad no deja de ser una similitud muy superficial. Para establecer similitudes reales debemos más bien buscar analogías en el funcionamiento matemático.

5. Universalidad de las computadoras digitales

Podemos situar las computadoras digitales que hemos tratado en el apartado anterior dentro de la categoría de «máquinas de estado discreto». Estas son máquinas que pasan mediante saltos o clics súbitos de un estado bastante definido a otro. Se trata de estados lo bastante distintos para que se dé la posibilidad de confusión entre ellos. Hablando en puridad no existen tales máquinas. En realidad, todo se mueve continuamente, pero podemos *considerar* positivamente muchos tipos de máquinas como de estado discreto. Por ejemplo, al referirnos a los interruptores de un sistema de iluminación, es una ficción convencional decir que cada uno de ellos debe hallarse totalmente conectado o desconectado. Pueden hallarse en posiciones intermedias,

pero en la mayoría de los casos podemos descartarlas. Como ejemplo de máquina de estado discreto consideremos una rueda que recorra 120° por segundo, pero que se detiene al accionar una palanca externa; ésta, además, en determinada posición, enciende una luz. Podríamos definir esta máquina de forma abstracta del siguiente modo: El estado interno de la máquina (descrito por la posición de la rueda) puede ser q_1 , q_2 o q_3 . Hay una señal de entrada i_0 o i_1 (posición de la palanca). El estado interno en cualquier momento está determinado por el último estado, y la señal de entrada lo estará con arreglo a la tabla

		Último estado		
		q_1	q_2	q_3
Entrada	i_0	q_2	q_3	q_1
	i_1	q_1	q_2	q_3

Las señales de salida, única indicación visible externa del estado interno (la luz), nos las da la tabla

Estado	q_1	q_2	q_3
Salida	o_0	o_0	o_1

Es un ejemplo clásico de máquina de estado discreto. Este tipo de máquinas se describen por medio de las tablas indicadas, a condición de que posean únicamente un número finito de estados posibles.

Podría parecer que, dado el estado inicial de la máquina y la señal de entrada, siempre fuera posible predecir los estados futuros, pero es una reminiscencia de la perspectiva de Laplace, según la cual, a partir del estado completo del universo en un momento del tiempo, definido por las posiciones y velocidades de todas sus partículas, se pueden predecir los estados futuros. Sin embargo, la predicción que estamos considerando es más próxima a la practicabilidad que la considerada por Laplace. El sistema del «universo como un todo» es de tal naturaleza que errores bastante pequeños en las condiciones iniciales pueden ejercer un efecto considerable en un momento futuro. El desplazamiento de un solo electrón en una billonésima de centímetro en un momento determinado puede ser la causa de que una persona muera aplastada por una avalancha un año más tarde o se libre de la catástrofe. Es una propiedad esencial de los sistemas mecánicos, que hemos denominado «máquinas de

estado discreto», el que semejante fenómeno no se produzca. Incluso si consideramos las actuales máquinas físicas en lugar de las máquinas idealizadas, el conocimiento razonablemente exacto de su estado en determinado momento nos procura un conocimiento razonablemente exacto de cualquier serie de pasos ulteriores.

Como hemos dicho, las computadoras digitales pertenecen al grupo de máquinas de estado discreto. Pero el número de estados que pueden adoptar este tipo de máquinas suele ser enormemente elevado. Por ejemplo, para la máquina que actualmente funciona en Manchester, la cifra aproximada sería de $2^{165.000}$, es decir de $10^{50.000}$ aproximadamente. Compárese esto con el citado ejemplo de la rueda que tenía tres estados. Se comprende sin dificultad por qué es tan elevado el número de estados. La computadora posee un almacenamiento correspondiente al papel que utiliza un computador humano. En este almacenamiento puede escribirse cualquiera de las combinaciones de símbolos que figurasen en el papel. Para simplificar, supongamos que sólo utilizamos como símbolos los dígitos del 0 al 9. No tomaremos en cuenta las variaciones de los signos manuscritos. Supongamos que la computadora dispone de 100 hojas de papel de 50 líneas cada una, con espacio para 30 dígitos. El número de estados será $10^{100 \times 50 \times 30}$, es decir, $10^{150.000}$. Esto equivale aproximadamente al número de estados de tres máquinas de Manchester juntas. El logaritmo con base dos del número de estados es en realidad lo que se denomina «capacidad de almacenamiento» de la máquina. Por lo tanto, la máquina de Manchester posee una capacidad de almacenamiento aproximada de 165.000, y la máquina con rueda del ejemplo mencionado, de aproximadamente 1-6. Si juntamos dos máquinas, habrá que sumar sus capacidades para saber la capacidad de la máquina resultante. Esto nos permite afirmar que «la máquina de Manchester contiene 64 pistas magnéticas, cada una de ellas con capacidad para 2.560, ocho tubos electrónicos con capacidad de 1.280. El almacenamiento diverso equivale aproximadamente a 300, lo que da un total de 174.380».

Disponiendo de la tabla correspondiente a una máquina de estado discreto se puede predecir lo que hará, y nada nos impide efectuar este cálculo con una computadora digital. A condición de que lo efectúe con suficiente rapidez, la computadora digital puede mimetizar el comportamiento de cualquier máquina de estado discreto. Entonces, se podría jugar con esa máquina (en el papel B) al juego de imitación y con la computadora digital mimetizante (en el papel de A), y el interrogador no sabría diferenciarlas. Naturalmente, la computadora digital debe poseer una capacidad de almacenamiento adecuada y funcionar a suficiente velocidad. Además, habrá que programarla expresamente para cada nueva máquina que se desee imitar.

Esta propiedad esencial de las computadoras digitales, por la que pueden imitar a cualquier máquina de estado discreto, se define diciendo que son máquinas *universales*. La existencia de máquinas con esta propiedad encierra la importante consecuencia de que, consideraciones de rapidez aparte, no hay necesidad de diseñar

diversas máquinas nuevas para que realicen los correspondientes nuevos procesos de computación. Todos pueden efectuarse con una sola computadora digital, convenientemente programada en cada caso. En consecuencia, como veremos, todas las computadoras digitales de este tipo son equivalentes en un sentido.

Ahora consideraremos la cuestión mencionada al final del apartado 3. Habíamos sugerido sustituir la pregunta «¿Pueden pensar las máquinas?» por la de «¿Existen computadoras digitales imaginables que jueguen bien al juego de imitación?» Si se desea, puede generalizarse más superficialmente esta pregunta: «¿Hay máquinas de estado discreto que hagan un buen juego?» Pero, dada la propiedad universal, vemos que ambas preguntas equivalen a: «Supongamos una determinada computadora digital C. ¿Es cierto que, modificando esta computadora para que tenga un almacenamiento adecuado y dotándola de un programa apropiado, podemos conseguir que C desempeñe eficazmente el papel de A en el juego de imitación y el papel de B lo haga un hombre?»

6. Opiniones contrapuestas sobre la cuestión principal

Consideremos ahora que hemos despejado el terreno y podemos ya pasar al debate de la pregunta «¿Pueden pensar las máquinas?» y de su variante, expuesta al final del apartado anterior. No podemos descartar totalmente la forma original del problema, ya que habrá diversidad de opiniones con respecto a la pertinencia de la sustitución y no podemos por menos que atender lo que se diga sobre el asunto.

Simplificaré las cosas para el lector si, en primer lugar, explico mi propia opinión sobre el tema. Consideremos primero la forma más exacta de la pregunta. Personalmente creo que, dentro de unos cincuenta años, se podrá perfectamente programar computadoras con una capacidad de almacenamiento aproximada de 10^9 para hacerlas jugar tan bien al juego de imitación que un preguntador corriente no dispondrá de más del 70 por ciento de las posibilidades para efectuar una identificación correcta a los cinco minutos de plantear las preguntas. Me parece que la pregunta original, «¿Pueden pensar las máquinas?», no merece discusión por carecer de sentido. No obstante, creo que, a finales del siglo, el sentido de las palabras y la opinión profesional habrán cambiado tanto que podrá hablarse de máquinas pensantes sin levantar controversias. Creo además que de nada sirve ocultar las ideas. La opinión tan generalizada de que los científicos proceden siempre de un hecho bien demostrado a otro hecho bien demostrado, y nunca se dejan influir por una conjetura no probada, es bastante errónea. A condición de que quede bien claro qué son hechos probados y qué son conjeturas, no existe ningún peligro. Las conjeturas son de suma importancia, porque sugieren posibles vías de investigación.

Ahora consideraré opiniones contrarias a la mía:

1. *La objeción teológica.* El pensamiento es una función del alma inmortal del hombre. Dios ha dado un alma inmortal a todos los hombres y mujeres, pero no a

ningún animal ni máquina. Por lo tanto, ni los animales ni las máquinas pueden pensar^[1].

Personalmente son ideas que rechazo totalmente, pero intentaré refutarlas en términos teológicos. La argumentación resultaría más convincente si se clasificara a los animales con el hombre, ya que existe mucha más diferencia, para mí, entre lo genuinamente animado y lo inanimado que entre el hombre y los animales. El carácter arbitrario de la opinión ortodoxa se evidencia aún más si tenemos en cuenta la opinión de los creyentes de otras religiones. ¿Cómo ve el cristianismo el dogma musulmán según el cual la mujer no tiene alma? Pero dejemos esto y volvamos a la cuestión principal. Creo que el citado argumento implica una grave restricción de la omnipotencia del Todopoderoso. Se admite así que hay cosas de las que El es incapaz, como es hacer que uno sea igual a dos, pero ¿dudaremos de su libertad para insuflar alma a un elefante, si a bien lo tiene? Cabe esperar que únicamente ejerciese tal poder en conjunción con una mutación que dotase al elefante de un cerebro mejorado que respondiera a las necesidades de esa alma. Podemos argüir exactamente lo mismo en el caso de las máquinas. Puede parecer distinto por ser más difícil de «tragar», pero esto únicamente significa que pensamos que es menos verosímil que El considere adecuadas las circunstancias para dotarlas de alma. Las circunstancias en cuestión se discuten en el resto de este trabajo. Al intentar construir este tipo de máquinas no estamos usurpando irreverentemente Su poder de crear almas, igual que no lo hacemos al procrear niños; en realidad, en ambos casos somos instrumentos de Su voluntad al procurar moradas para las almas que El crea.

Pero todo esto es mera especulación. No me impresionan mucho los argumentos teológicos, aunque se utilicen como apoyo. A lo largo de la historia se ha comprobado cuánto dejan que desear. En tiempos de Galileo se argumentaba que las Sagradas Escrituras decían: «Y el sol se detuvo... y no fue hacia el ocaso durante casi un día» (Josué x. 13) y que: «El creó los fundamentos de la Tierra para que no se moviera» (Salmo cv. 5) como refutación convincente de la teoría copernicana. Con los conocimientos actuales estos argumentos resultan fútiles, pero en una época de escasos conocimientos científicos causaban muy distinta impresión.

2. *La objeción del «avestruz».* «Las consecuencias de que las máquinas piensen serían horribles. Creamos y esperemos que no sea posible».

Este argumento rara vez se expone de forma tan abierta, pero afecta a la mayoría de quienes reflexionamos sobre ello. Nos gusta creer que el hombre es en algún modo superior al resto de la creación, y tanto mejor si podemos demostrar que es *necesariamente* superior, pues entonces no existe peligro de que pierda su posición dominante. La popularidad del argumento teológico está claramente vinculada a esta idea y cuenta con muchos adeptos entre los intelectuales, pues éstos aprecian más que otras personas el poder del pensamiento y se muestran más inclinados a basar su convencimiento de la superioridad del hombre en este poder.

No creo que este argumento sea lo bastante fundado para molestarme en refutarlo. Tal vez sea mejor consolarse, buscándolo quizás en la transmigración de las almas.

3. *La objeción matemática.* Pueden citarse toda una serie de resultados de la lógica matemática para demostrar que hay limitaciones en el poder de las máquinas de estado discreto. El más conocido es el denominado teorema de Gödel, que demuestra que en cualquier sistema lógico lo bastante potente pueden formularse afirmaciones que no pueden demostrarse ni refutarse dentro del sistema, salvo en caso de que posiblemente tal sistema sea incoherente. En ciertos aspectos similares hay otros resultados expuestos por Church^[*], Kleene, Rosser y Turing. La tesis de este último autor es la que merece mayor consideración en este caso, por referirse específicamente a las máquinas, mientras que las de los otros sólo son utilizables en tanto que argumentos relativamente indirectos: si, por ejemplo, recurrimos al teorema de Gödel, hace falta a la vez disponer de medios para describir los sistemas lógicos en términos de máquinas y las máquinas en términos de sistemas lógicos. El resultados en cuestión se refiere a un tipo de máquina que es fundamentalmente una computadora digital con capacidad infinita, y postula que hay ciertas cosas que esa máquina no puede efectuar. Si se la equipa para dar respuesta a preguntas como en el juego de imitación, habrá preguntas que contestará mal o no podrá contestar por mucho tiempo que se le conceda. Naturalmente, puede haber muchas preguntas de esta clase, y preguntas que no pueda contestar satisfactoriamente una máquina las contestará adecuadamente otra. Desde luego estamos por ahora en la suposición de que estas preguntas son de tal índole que la respuesta es «Sí» o «No», y no preguntas del tipo «¿Qué opinas sobre Picasso?» Las preguntas que sabemos que la máquina no contesta son de esta clase: «Supongamos una máquina con las siguientes características... ¿contestará esta máquina 'Sí' a cualquier pregunta?» Los puntos suspensivos se sustituyen por la descripción de una máquina modelo estándar, que podría ser como la que se cita en el apartado 5. Si la máquina descrita guarda cierta relación comparativamente simple con la máquina a que se está interrogando, puede demostrarse que la respuesta es incorrecta o no se va a producir. Este es el resultado matemático: se arguye que demuestra una incapacidad por parte de las máquinas a la que no está expuesto el intelecto humano.

La respuesta taxativa a este razonamiento es que, aunque está demostrado que existen limitaciones en la capacidad de cualquier máquina, sólo se ha afirmado, sin prueba alguna, que tales limitaciones no son aplicables al intelecto humano. Sin embargo, yo no creo que esta posibilidad pueda rechazarse tan alegremente. Cuando se plantea a una de estas máquinas la pregunta crítica adecuada y nos da una respuesta concreta, sabemos que la respuesta es incorrecta y esto nos da cierta sensación de superioridad. ¿Es una sensación ilusoria? Sin duda es lo bastante legítima, pero yo no creo que haya que atribuirle demasiada importancia. También nosotros en muchas ocasiones respondemos erróneamente a preguntas, lo cual no justifica esa enorme sensación de halago al ver que las máquinas fallan. Además, sólo

podemos sentir en este caso nuestra superioridad en relación con la máquina concreta, objeto de nuestra frágil victoria. No es un triunfo simultáneo frente a *todas* las máquinas. En resumen, habrá hombres más listos que cualquier máquina, pero también otras máquinas más listas, y así sucesivamente.

Los partidarios del argumento matemático aceptarán en su mayoría —creo yo— que el juego de imitación es una buena base para la discusión. A los partidarios de las dos primeras objeciones seguramente no les interesará ningún razonamiento.

4. *El argumento de la conciencia.* Este argumento está perfectamente expresado en un discurso conmemorativo del profesor Jefferson, en 1949, del que cito: «Hasta que una máquina sea capaz de escribir un soneto o de componer un concierto, porque tenga la facultad de reflexionar y sea capaz de sentir, y no por la combinación aleatoria de símbolos, no podremos admitir que esa máquina sea igual al cerebro, en el sentido de que no sólo los escriba, sino que sepa que los ha escrito. Ningún mecanismo (y no hablo de una señal artificial, invención simplona) puede sentir placer por sus logros, pena cuando se funden sus válvulas, regocijo por los halagos, depresión por sus errores, atracción sexual, enfado o decepción cuando no consigue lo que quiere».

Este argumento parece ser la negación de la validez de nuestro test. Según la modalidad más extremada de este tipo de planteamiento, la única manera de asegurarse de que una máquina piensa es *ser* la máquina y sentir el propio pensamiento. Sólo entonces pueden exponerse tales sentimientos a todo el mundo, pero tampoco está justificado que a nadie le importen. Según este planteamiento, también la única manera de saber que una *persona* piensa es ser esa persona concreta. De hecho, es un punto de vista solipsista. Puede que sea el punto de vista más lógico, pero dificulta la comunicación de ideas. A puede sentirse inclinado a creer «A piensa pero B no», mientras que B creerá que «B piensa pero A no». En lugar de discutir indefinidamente este punto, mejor es adscribirse al cortés convencionalismo de que todos piensan.

Estoy convencido de que el profesor Jefferson no desea adoptar el punto de vista extremo y solipsista. Probablemente se halle dispuesto a aceptar como prueba el juego de imitación. El test (omitiendo el jugador B) suele usarse en la práctica bajo la denominación de *examen oral* para descubrir si el candidato entiende de verdad algo o lo «ha aprendido como un papagayo». Escuchemos un extracto de uno de esos *exámenes orales*:

Examinador: En el primer verso de su soneto, que dice «¿Te compararía con un día de verano?», ¿no sería igual, o mejor, «un día de primavera»?

Examinado: No rimaría.

Examinador: ¿Y «un día de invierno»? Rima perfectamente.

Examinado: Sí, pero a nadie le gusta que le comparen con un día de invierno.

Examinador: ¿Diría usted que Mr. Pickwick le recuerda la Navidad?

Examinado: En cierto modo.

Examinador: Pues Navidad es un día de invierno, y no creo que a Mr. Pickwick le molestara la comparación.

Examinado: Creo que bromea usted. Por día de invierno se entiende un día de invierno genuino y no uno especial como el de Navidad.

Y así sucesivamente. ¿Qué diría el profesor Jefferson si la máquina escritora de sonetos fuera capaz de contestar así en el examen oral? No sé si la consideraría accionada por «una simple señal artificial» al dar tales respuestas, pero, si las respuestas fueran tan adecuadas y coherentes como en el párrafo anterior, no creo que las calificara de «invención simplona». Yo creo que con esta expresión se intenta definir dispositivos tales como la inclusión en la máquina de un disco de alguien que lee un soneto, dotado del correspondiente relé que lo conecte de vez en cuando.

En resumen, creo que a la mayoría de los partidarios del argumento de la conciencia se les podría convencer de que lo abandonarían en lugar de forzarles a la actitud solipsista. Entonces, probablemente se inclinasen a aceptar la prueba.

No quisiera dar la impresión de que creo que no existe misterio en lo que se refiere a la conciencia. Existe, por ejemplo, algo así como una paradoja en relación con su localización. Pero no creo que haya que solucionar necesariamente ese misterio para responder a la cuestión que nos ocupa en este trabajo.

5. *Argumentos de incapacidades diversas.* Estos argumentos responden al esquema: «Te aseguro que pueden hacerse máquinas que realicen todo lo que has dicho, pero es imposible construir una máquina que haga X», y se citan al respecto diversas X. A continuación expongo una selección:

«Ser amable, ingeniosa, hermosa, amistosa» (pág. 33), «poseer iniciativa, tener sentido del humor, distinguir entre lo bueno y lo malo, cometer faltas» (pág. 33), «enamorarse, apreciar las fresas y los helados» (pág. 33), «enamorar a alguien, aprender por la experiencia» (págs. 41 y ss.), «utilizar adecuadamente las palabras, ser objeto de su propio pensamiento» (pág. 34), «tener un comportamiento tan versátil como una persona, hacer algo auténticamente nuevo» (pág. 35). (Algunas de estas facultades parecen recibir mayor atención al citarse un número de página a continuación).

Generalmente estas afirmaciones no se apoyan en razonamientos y, personalmente, creo que en esencia se basan en el principio de la inducción científica. Una persona ve miles de máquinas durante su vida y, por lo que ve de ellas, extrae una serie de conclusiones generales. Son feas y cada una de ellas está ideada para una tarea concreta; cuando se desea que ejecuten varias funciones, son inservibles, su variedad de comportamiento es muy limitada, etc., etc. En consecuencia, concluye que ésas son las características de las máquinas en general. Muchas de estas

limitaciones se asocian a la escasa capacidad de almacenamiento de la mayoría de las máquinas (supongo que, en el concepto de capacidad de almacenamiento, se incluyen en cierto modo a las máquinas distintas a las de estado discreto. No importa la definición exacta, ya que no aspiramos a una exactitud matemática en esta discusión). Hace unos años, cuando aún se hablaba poco de computadoras digitales, era de esperar que su mención suscitara incredulidad cuando se hablaba de sus propiedades sin explicar su construcción. Supongo que era también debido a la aplicación del principio de inducción científica. Naturalmente esta clase de aplicación del principio suele ser inconsciente. Cuando un niño que ha sufrido una quemadura teme al fuego y demuestra que lo teme evitándolo, decimos que está aplicando la inducción científica. (Naturalmente, puedo también describir su comportamiento de muchas otras maneras.) Los trabajos y las costumbres humanos no parecen constituir un material muy adecuado para la aplicación de la inducción científica. Habría que investigar una gran magnitud espacio-temporal para obtener resultados fiables, pues, si no, creeremos (como la inmensa mayoría de los niños ingleses) que todo el mundo habla inglés y que es una tontería aprender francés.

Sin embargo, conviene hacer algunas observaciones respecto de las múltiples incapacidades que hemos citado. La incapacidad para apreciar las fresas y los helados le habrá parecido al lector una futilidad. Puede que se construya una máquina que aprecie esos manjares, pero sería una imbecilidad intentarlo. Lo importante respecto de esta incapacidad es que está destinada a aumentar el número de incapacidades, por ejemplo, el mismo tipo de dificultad de comunicación amistosa que se produce entre el hombre y la máquina también se da entre un hombre blanco y otro hombre blanco, o entre un hombre negro y otro hombre negro.

Afirmar que las «máquinas no cometen errores» parece curioso. Se siente uno inclinado a replicar: «¿Y son por eso peores?», pero adoptemos una actitud más simpática y tratemos de comprender qué es lo que significa. Creo que esta crítica puede explicarse en términos del juego de imitación. Se afirma que al preguntador le basta, para distinguir una máquina del hombre, plantear una serie de problemas aritméticos. La máquina queda desenmascarada por su tremenda exactitud. Así de sencillo, pero la máquina (programada para jugar el juego) no tratará de dar las respuestas *correctas* a los problemas aritméticos e introducirá deliberadamente errores de modo calculado para confundir al preguntador. Una avería mecánica se percibirá probablemente al darse una decisión inadecuada respecto del tipo de error aritmético a efectuar. Incluso esta interpretación crítica no es lo bastante simpática, pero no disponemos de espacio para extendernos más. A mí me parece que la crítica se fundamenta en una confusión de dos tipos de error. Podemos denominarlos «errores de funcionamiento» y «errores de conclusión». Los errores de funcionamiento los causa un efecto mecánico o eléctrico que obliga a la máquina a comportarse de modo distinto a como está diseñada. En las discusiones filosóficas se ignora la posibilidad de tales errores y se habla de «máquinas abstractas». Estas

máquinas abstractas son ficciones matemáticas más que objetos físicos. Son, por definición, incapaces de errores de funcionamiento. En este sentido podemos afirmar con certeza que «las máquinas no cometen errores». Los errores de conclusión sólo pueden producirse cuando se atribuye un significado a las señales de salida de la máquina. La máquina puede, por ejemplo, imprimir ecuaciones matemáticas, o frases en inglés. Cuando escribe una oración incorrecta, decimos que ha cometido un error de conclusión. Evidentemente no existe motivo para decir que una máquina no puede cometer este tipo de error. Puede que se limite a escribir sin parar « $0 = 1$ ». Adoptando un ejemplo menos peyorativo, digamos que, al estar dotada de un método para extraer conclusiones por inducción científica, es presumible que semejante método conduzca a veces a resultados erróneos.

A la afirmación de que una máquina no puede ser objeto de su propio pensamiento sólo puede contestarse si se demuestra que la máquina posee algún pensamiento referido a *algún* tema. No obstante, «el tema de las operaciones de una máquina» parece significar algo, al menos para quienes trabajan con ella. Si, por ejemplo, la máquina trata de hallar la solución a la ecuación $x^2 - 40x - 11 = 0$, uno no puede resistir la tentación de calificar esta ecuación de objeto parcial del tema de la máquina en ese momento. En este aspecto no cabe duda de que una máquina es su propio objeto, ya que se la puede utilizar para que contribuya a la confección de su propio programa, o para predecir el efecto de alteraciones en su propia estructura. Observando los resultados de su propio comportamiento, es capaz de modificar sus programas para efectuar determinada tarea con mayor eficacia. Son posibilidades de un futuro no muy lejano, no sueños utópicos.

La crítica de que una máquina no puede tener versatilidad de comportamiento es sólo una manera de decir que no puede tener una gran capacidad de almacenamiento. Hasta hace relativamente poco tiempo una simple capacidad de mil dígitos era algo extraordinario.

Las críticas que estamos considerando suelen ser variantes enmascaradas del argumento de la conciencia. Generalmente, si uno sostiene que una máquina *puede* hacer una de esas cosas y describe la clase de método del que puede servirse, no se logra impresionar a los detractores, pues piensan que el método (sea el que fuere, por ser mecánico necesariamente) es algo vil. Cotéjese el paréntesis del párrafo de Jefferson citado anteriormente.

6. *Objeción de lady Lovelace*. La información más pormenorizada sobre la máquina analítica de Babbage figura en un informe de lady Lovelace. En él se afirma: «La Máquina Analítica no pretende *crear* nada. Puede realizar *lo que nosotros sepamos mandarle*» (en cursiva en el informe original). Es Hartree quien cita este párrafo, y añade: «Esto no implica que sea imposible construir equipo electrónico que “piense por sí solo”, o en el que, en términos biológicos, no se pueda implantar un reflejo condicionado que sirva de base al “aprendizaje”. Si es o no posible en principio, es una cuestión apasionante y estimulante, esbozada en algunos de los

últimos avances tecnológicos. Pero no parecía que las máquinas construidas en aquella época tuvieran tal propiedad».

Coincido totalmente con Hartree al respecto. Advuértase que él no afirma que la máquina en cuestión no posea la propiedad, sino que a lady Lovelace no le constaba que la tuviera. Es muy posible que las máquinas en cuestión tuvieran en cierto modo esa propiedad. Supongamos que una máquina de estado discreto tiene esa propiedad. La Máquina Analítica era una computadora digital universal, de forma que, si su capacidad de almacenamiento y su velocidad eran adecuados, con un programa idóneo se la podría inducir a mimetizar la propia máquina. Probablemente este razonamiento no se le ocurrió a la condesa ni al propio Babbage. En cualquier caso, ellos no tenían por qué reivindicar todo lo reivindicable.

Volveremos a hablar del tema en el apartado de máquinas que aprenden.

Una variante a la objeción de lady Lovelace afirma que las máquinas «nunca hacen nada nuevo». Podemos parangonar tal afirmación al refrán: «No hay nada nuevo bajo el sol». ¿Quién puede tener el firme convencimiento de que el «trabajo original» que se acaba de realizar no es sino el desarrollo de la simiente que ha dejado en él el aprendizaje, o la consecuencia de atenerse a consabidos principios generales? Otra variante mejor de esta objeción es la de que la máquina nunca «puede sorprendernos». Es un desplante más directo, por lo que respondemos directamente. Las máquinas me sorprenden muy a menudo. Fundamentalmente porque no calculo lo suficiente para figurarme lo que van a hacer, o, más bien, porque, aunque calculo, lo hago de forma precipitada, descuidada y corriendo riesgos, y me digo: «Supongo que el voltaje es aquí el mismo que allí; bueno, supongamos que es el mismo». Naturalmente, muchas veces me equivoco, el resultado me sorprende, aunque, una vez finalizado el experimento, me olvido de mis falsas suposiciones. Con esta confesión me expongo a sermones sobre mis malas costumbres, pero no empañó mi sinceridad al dar fe de las sorpresas que experimento.

No pretendo con esta réplica silenciar la crítica. Probablemente puede deducirse que tales sorpresas se deben a algún acto creativo mental por mi parte, y nada dicen a favor de la máquina. Esto nos obliga a volver al argumento de la conciencia, muy lejos de la idea de sorpresa. Es un tipo de argumentación muy similar, pero quizá valga la pena señalar que la apreciación de algo como sorprendente requiere igual «acto mental creativo», independientemente de que la sorpresa la cause una persona, un libro, una máquina o lo que sea.

La opinión de que las máquinas no pueden producir sorpresa se basa, creo yo, en el sofisma en el que suelen incurrir particularmente filósofos y matemáticos: la asunción de que, cuando a la mente se le presenta un hecho, todas las consecuencias del mismo la invaden con él simultáneamente. Es una asunción muy útil en muchas circunstancias, pero se olvida con harta facilidad de que es falsa. Una consecuencia natural de asumirla como cierta es que se da por sentado que no hay mérito en la simple elucidación de consecuencias a partir de datos y principios generales.

7. *Argumento de la continuidad del sistema nervioso.* Desde luego el sistema nervioso no es una máquina de estado discreto. Un pequeño error de información sobre la magnitud de un impulso nervioso aferente en una neurona puede modificar considerablemente la magnitud del impulso de salida. Puede argüirse que, precisamente por eso, no cabe posibilidad de mimetizar el comportamiento del sistema nervioso mediante un sistema de estado discreto.

Cierto es que una máquina de estado discreto es distinta a una máquina continua, pero, si nos ceñimos a las condiciones del juego de imitación, el preguntador no gana nada con esa diferencia. Podemos aclarar la situación si consideramos cualquier otra máquina continua más sencilla. Un analizador diferencial, por ejemplo. (Un analizador diferencial es un tipo de máquina de estado no discreto que se emplea para cierta clase de cálculos). Algunos dan la respuesta impresa, por lo que son adecuados para intervenir en el juego. Una computadora digital no puede predecir exactamente las respuestas que da a un problema el analizador diferencial, pero sí puede dar la respuesta correcta. Por ejemplo, si se pregunta el valor de π (3,1416 aproximadamente), es razonable elegir al azar entre los valores 3,12 3,13, 3,14, 3, 15, 3,16 con las probabilidades de 0,05, 0,15, 0,55, 0, 19, 0,06 (pongamos por caso). En tales circunstancias resultará muy difícil para el preguntador distinguir al analizador diferencial de la computadora digital.

8. *El argumento de la informalidad de comportamiento.* No se puede elaborar un conjunto de reglas para describir lo que una persona hace en todas las circunstancias concebibles. Puede establecerse la regla de que, por ejemplo, hay que detenerse al ver un semáforo rojo y continuar si se ve uno verde, pero ¿qué sucede si, por un error, se iluminan los dos a la vez? Quizá la persona decida que es mejor detenerse. Pero por esta decisión pueden surgir ulteriormente dificultades. Intentar sentar reglas de conducta que cubran cualquier eventualidad, hasta las resultantes de las luces de tráfico, parece imposible. Estoy de acuerdo con esto.

A partir de ello se arguye que no podemos ser máquinas. Trataré de exponer el argumento, pero temo no hacerle debidamente justicia. Al parecer, se desarrolla de este modo: «Si cada persona posee un conjunto fijo de reglas de conducta por las que rige su vida, no sería más que una máquina; pero no hay tales reglas. Por lo tanto, las personas no pueden ser máquinas». Es deslumbrante el injusto medio. No creo que el argumento se plantee casi nunca así, pero estoy convencido de que constituye la base de la argumentación. Sin embargo, puede darse cierta confusión entre «reglas de conducta» y «leyes de comportamiento» para oscurecer la conclusión. Por «reglas de conducta» entiendo preceptos tales como «Pare si ve luces rojas» que uno puede cumplir conscientemente. Por «leyes de comportamiento» entiendo leyes naturales aplicables al cuerpo humano, tales como «si le pellizcas, chilla». Si sustituimos «leyes de comportamiento que regulan su vida» por «leyes de conducta por las que rige su vida», el injusto medio deja de ser insuperable en el argumento en cuestión, pues creemos que no sólo es cierto que estar regulado por leyes de comportamiento

implica ser una especie de máquina (aunque no necesariamente una máquina de estado discreto), sino que, a la inversa, ser tal máquina implica estar regulado por tales leyes. Sin embargo, no podemos convencernos tan fácilmente de la ausencia total de leyes de comportamiento como de la ausencia absoluta de leyes de conducta. El único modo de descubrir tales leyes consiste en la observación científica, y no conocemos circunstancias en las que pueda decirse: «Ya hemos buscado bastante. No existen tales leyes».

Podemos demostrar más categóricamente que semejante afirmación es injustificada. Supongamos que fuera posible con absoluta seguridad descubrir esas leyes, si existiesen. Entonces, dada una máquina de estado discreto, no cabe duda de que podría descubrirse, mediante la observación suficiente para predecirlas, su comportamiento futuro, y eso dentro de un tiempo razonable, digamos mil años. Pero no parece ser el caso. He elaborado en la computadora de Manchester un pequeño programa con tan sólo 1.000 unidades de almacenamiento, merced al cual, si se entrega a la máquina una cifra de dieciséis guarismos, responde con otra de igual magnitud en dos segundos. Desafío a cualquiera a que descubra en esas respuestas suficientes datos sobre el programa para ser capaz de predecir cualquier respuesta a valores no probados.

9. *El argumento de la percepción extra-sensorial.* Supongo que el lector está al corriente de la idea de percepción extra-sensorial y del significado de sus cuatro variantes: telepatía, clarividencia, precognición y psicocinesis. Estos extraños fenómenos parecen refutar todas las ideas cinéticas habituales, ¡Cuánto nos gustaría desacreditarlos! Pero lamentablemente la evidencia estadística, al menos en el caso de la telepatía, es abrumadora. Resulta difícil para cualquiera reajustar sus propias ideas para dar cabida a estos hechos singulares, pero, una vez admitidos, no parece que cueste mucho creer en fantasmas y espíritus. Lo primero que se nos ocurre es la idea de que nuestros cuerpos se mueven de modo simple con arreglo a las leyes físicas conocidas, junto a otras no descubiertas pero bastante parecidas.

Para mí es un argumento de bastante peso. Podría argüirse que muchas teorías científicas siguen siendo válidas en la práctica, a pesar de que contradigan la percepción extra-sensorial, y que puede prescindirse perfectamente de ella, pero no deja de ser un conformismo fácil; precisamente es muy de temer que no sea el pensamiento el tipo de fenómeno en el que la percepción extra-sensorial sea particularmente relevante.

Un argumento más específico basado en la percepción extra-sensorial sería el siguiente: «Juguemos al juego de imitación, teniendo por testigo a una persona que sea buena receptora telepática y a una computadora digital. El preguntador puede plantear preguntas de este tipo: “¿A qué palo pertenece la carta que tengo en mi mano derecha?” La persona, mediante telepatía o clarividencia, da la respuesta correcta 130 veces sobre 400 cartas. La máquina sólo puede adivinar al azar y tal vez acierte 104 veces, y así el preguntador efectúa la identificación correcta». Esta es una interesante

posibilidad. Supongamos que la computadora digital cuenta con un generador numérico aleatorio, es natural que lo utilice para dar la respuesta. Pero el generador numérico aleatorio está sujeto al poder psicocinético del preguntador y quizás esta psicocinesis sea la causa de que la máquina acierte más veces de las que cabría esperarse de un cálculo de probabilidades, por lo que el preguntador seguiría siendo incapaz de efectuar la identificación correcta. Por otra parte, puede ser capaz de acertar sin plantear preguntas, gracias a la clarividencia. Con la percepción extra-sensorial puede suceder cualquier cosa.

Si admitimos la telepatía, habrá que depurar la prueba. Puede considerarse la situación similar a la que se produce si el preguntador hablara consigo mismo y uno de los participantes estuviera escuchando con el oído en la pared. Situando a los participantes en una «habitación a prueba de telepatía», se restablecerían las condiciones.

7. Máquinas que aprenden

Habría comprobado el lector que no dispongo de argumento positivo alguno lo bastante convincente para apoyar mi tesis. Si lo tuviera, no me habría tomado tanta molestia en exponer detalladamente las falacias de las tesis contrarias. Ahora expondré la evidencia en favor de mi punto de vista.

Volvamos brevemente a la objeción de lady Lovelace, quien afirmaba que la máquina sólo puede hacer lo que nosotros le mandemos. Podríamos decir que una persona puede «inyectar» una idea en una máquina y que ésta respondería hasta cierto límite, quedándose quieta a continuación, como la cuerda de un piano percutida por un martillo. Otro símil podría ser una pila atómica de tamaño inferior al crítico: una idea inyectada correspondería a un neutrón que penetra desde fuera en la pila. Cada uno de estos neutrones provoca una determinada alteración que acaba por disiparse. Sin embargo, si aumentamos suficientemente el tamaño de la pila, la alteración causada por el neutrón incluso irá en aumento hasta la completa destrucción de la pila. ¿Existe un fenómeno equivalente para las mentes, y se da también en el caso de las máquinas? En el caso de la mente humana parece haberlo. La mayoría de los cerebros parecen ser «subcríticos», es decir, que corresponden en esta analogía a pilas de tamaño subcrítico. Una idea presentada a este tipo de mente, no inducirá generalmente más que una idea por respuesta. Una reducidísima proporción de cerebros son supercríticos. En ellos una idea da origen a toda una «teoría» formada por ideas secundarias, terciarias y de todo orden. Las mentes animales parecen decididamente ser subcríticas. Siguiendo la analogía, nos preguntamos: «¿Se puede hacer que una máquina sea supercrítica?»

La analogía de la «piel de cebolla» también es válida. Si consideramos las funciones de la mente o del cerebro, observamos determinadas operaciones explicables en términos puramente mecánicos. Lo que decimos no es aplicable a la

auténtica mente: es una especie de piel que hay que quitar si queremos verla realmente. Pero, luego, en lo que queda, encontramos otra piel que hay que eliminar, y así sucesivamente. Con este método, ¿llegamos con seguridad a la mente «real», o simplemente a la piel que no encierra nada? En tal caso toda mente es mecánica. (De todas formas, hemos explicado ya que no es una máquina de estado discreto.)

Los últimos párrafos no pretenden ser argumentos convincentes, sino más bien deben tomarse como una letanía destinada a inculcar una creencia».

El único apoyo realmente satisfactorio que puede darse a la opinión manifestada al principio del apartado 6 es el que consiste en esperar a finales de siglo y luego efectuar el experimento señalado. ¿Pero qué podemos decir entretanto? ¿Qué pasos hemos de dar ahora para que dé buen resultado el experimento?

Como he dicho, el problema fundamental estriba en programar. También serán imprescindibles progresos de ingeniería, pero creo que estarán a la altura de las necesidades. Las estimaciones de la capacidad de almacenamiento del cerebro oscilan entre 10^{10} y 10^{15} dígitos binarios. Personalmente me inclino por el valor más bajo y creo que sólo una pequeña parte se utiliza para los tipos más elevados de pensamiento. La mayor parte de esta capacidad se emplea seguramente en la retención de impresiones visuales. Me sorprendería que se necesitara más de 10^9 para jugar bien al juego de imitación, en cualquier caso contra un hombre ciego. (Nota: la capacidad de la *Encyclopaedia Britannica*, décimo primera edición, es de 2×10^9). Una capacidad de almacenamiento de 10^7 sería una posibilidad bastante real, aun con las técnicas actuales. Probablemente no será preciso aumentar la velocidad de operación de las máquinas. Partes de las máquinas modernas, que podríamos calificar de auténticas células nerviosas, trabajan mil veces más rápido que éstas. Con esto se conseguiría un «margen de seguridad» para compensar pérdidas de velocidad producidas por diversos motivos. El problema estriba, en último extremo, en saber cómo programar estas máquinas para jugar al juego. Con mi actual ritmo de trabajo produzco unos mil dígitos de programa diarios; en consecuencia, unas sesenta personas, trabajando asiduamente durante cincuenta años, podrían llevar a cabo esta tarea si no traspapelaran nada. Parece deseable un método más expeditivo.

En el proceso de intentar la imitación de una mente humana adulta estamos obligados a pensar muy en serio sobre el proceso por el que se ha llegado al estado en que se halla. Se observarán tres factores:

- a. El estado inicial de la mente al nacer.
- b. La educación que ha tenido.
- c. Otras experiencias, aparte de la educación, a que haya estado sometida.

En lugar de intentar la elaboración de un programa que imite la mente adulta, ¿por qué no establecer uno que simule la mente infantil? Si luego la sometemos a un curso adecuado de formación, podría obtenerse un cerebro adulto. Podemos decir que el cerebro infantil es como el cuaderno recién comprado en una papelería: poco

mecanismo y muchas hojas en blanco. (Mecanismo y escritura son casi sinónimos desde nuestro punto de vista). Nuestra esperanza se funda en que hay tan poco mecanismo en el cerebro infantil que debe resultar fácil programar algo similar. Podemos suponer que la cantidad de trabajo formativo, en una primera aproximación, sea muy parecida a la aplicable en el caso de un niño.

De este modo, el problema queda dividido en dos partes: el programa infantil y el proceso formativo. Ambos estrechamente vinculados. No puede esperarse construir una buena máquina infantil al primer intento; hay que experimentar enseñando a la máquina, y comprobar si aprende bien. Luego puede probarse otra vez y ver si es mejor o peor. Evidentemente existe una clara relación por analogía entre este proceso y el de la evolución:

Estructura de la máquina infantil = Material hereditario

Cambios de la máquina infantil = Mutaciones

Selección natural = Juicio del experimentador

Sin embargo, es de esperar que este proceso sea más expeditivo que el de la evolución. La supervivencia del más apto es un método lento para valorar las ventajas. El experimentador, aplicando su inteligencia, debe ser capaz de acelerarlo. De igual importancia es el hecho de que no está limitado por mutaciones aleatorias. Si el experimentador descubre la causa de determinada debilidad, puede probablemente decidir el tipo de mutación que la mejore.

A la máquina no se le podrá aplicar exactamente el mismo proceso de aprendizaje que a un niño. Ya que, por ejemplo, no tendrá piernas y no se le podrá ordenar que vaya a por un cubo de carbón. Seguramente tampoco tendrá ojos. Y por mucho que se compensen estas deficiencias con una buena ingeniería, no se podrá enviar a la criatura a la escuela porque sería motivo de burla de sus compañeros. Habrá que darle clases particulares, sin preocuparnos por las piernas, los ojos, etc. El caso de Helen Keller^[*] demuestra que es posible la labor educativa a condición de que se establezca una comunicación bilateral entre maestro y alumno por el medio que sea.

Normalmente asociamos castigos y recompensas al proceso educativo. Algunas máquinas infantiles simples pueden construirse o programarse ateniéndose a ese principio. Hay que construir la máquina de tal modo que los acontecimientos que preceden brevemente a la aparición de la señal de castigo cuenten con mínimas posibilidades de repetición, y que, por el contrario, la señal de recompensa incremente la posibilidad de repetición de secuencias que la motivan. Estas especificaciones no presuponen tipo de sentimiento alguno por parte de la máquina. He realizado algunos experimentos con este tipo de máquina infantil y he logrado enseñarle varias cosas, pero utilicé un método de aprendizaje excesivamente heterodoxo para que el experimento pueda considerarse un éxito.

El empleo de castigos y recompensas puede a lo sumo formar parte del proceso de aprendizaje. En términos generales, si el enseñante no dispone de otros medios de comunicación con el alumno, la cantidad de información que éste recibe nunca excede el número de recompensas y castigos. Cuando un niño ha aprendido finalmente a repetir «Casabianca», se sentirá probablemente muy afligido si la única manera de dilucidar el texto es la técnica de las «Veinte preguntas» y cada «NO» supone una bofetada. Por lo tanto, es necesario disponer de otros canales de comunicación «no emocionales». Si los hay, se puede enseñar a una máquina por el método de premios y castigos a obedecer órdenes dadas en una lengua determinada, es decir un lenguaje simbólico. Estas órdenes se transmiten por canales «no emocionales», y el empleo de dicho lenguaje disminuye notablemente la cantidad de castigos y premios.

Puede existir diversidad de opiniones en cuanto a la complejidad adecuada de la máquina infantil. Puede intentarse una construcción lo más simple posible, coherente con los principios generales. O puede dotársela de un sistema completo integrado^[2] de inferencia lógica, en cuyo caso el almacenamiento estará fundamentalmente ocupado por definiciones y proposiciones. Estas proposiciones serían de diversa índole: hechos bien establecidos, conjeturas, teoremas matemáticamente demostrables, afirmaciones hechas por una autoridad, expresiones con forma lógica de proposición pero de valor no creíble. Algunas proposiciones serían «imperativas». La máquina estaría construida de forma que, en cuanto una imperativa se clasificara como «bien establecida», se produjera automáticamente la acción apropiada. Como ejemplo, supongamos que el maestro dice a la máquina: «Ahora haz los deberes». Esto podría dar lugar a que «El maestro dice “Ahora haz los deberes”» quedara incluido en los hechos bien establecidos. Otra posibilidad sería: «Todo lo que dice el maestro es cierto». Ambas posibilidades combinadas podrían dar por resultado que la imperativa «Ahora haz los deberes» quedara incluida entre los hechos bien establecidos, lo cual, con arreglo a la construcción de la máquina, significaría que se inician realmente los deberes, pero el efecto es muy poco satisfactorio. El proceso de inferencia que utilice la máquina tiene que satisfacer al lógico más riguroso. Por ejemplo, no habrá jerarquía de tipos, lo que no significa que no se produzcan falacias de tipos, semejantes al riesgo de caer por un precipicio no señalizado. Unos imperativos adecuados (expresados *dentro* de los sistemas, pero que no formen parte de las reglas *del* sistema), tales como «No emplees una clase si no es una subclase de las mencionadas por el maestro», ejercerían la misma función que un letrado que indicara: «No acercarse al borde».

Las imperativas a las que obedece una máquina sin miembros son necesariamente de índole intelectual, como en el ejemplo citado (hacer los deberes). Entre dichas imperativas son importantes las que rigen el orden en que hay que aplicar las reglas del sistema lógico correspondiente, ya que, en cada fase de la utilización de un sistema lógico, hay una amplia alternativa de pasos que pueden seguirse para no

transgredir las reglas de ese sistema lógico. Estas opciones marcan la diferencia entre un razonador brillante y otro torpe, pero no la diferencia entre uno serio y otro tramposo. Las proposiciones que conducen a las imperativas de esta clase pueden ser: «Cuando se mencione a Sócrates, utiliza el silogismo en Bárbara», o «Si se ha demostrado que un método es más rápido que otro, no uses el método lento». Algunas pueden «basarse en una autoridad», pero otras puede producirlas la propia máquina por inducción científica, por ejemplo.

La idea de una máquina que aprende puede parecer paradójica a algunos lectores. ¿Cómo pueden cambiarse las reglas de operación de la máquina? Estas deben especificar punto por punto cómo debe reaccionar la máquina independientemente de su historia y al margen de los cambios que experimente. Por lo tanto, las reglas son bastante invariables con respecto al tiempo. Y es bien cierto. La explicación de la paradoja consiste en que las reglas que cambian en el proceso de aprendizaje son de un tipo menos pretencioso y sólo tienen validez efímera. El lector puede establecer un paralelismo con la Constitución de los Estados Unidos.

Una característica importante de la máquina que aprende es la de que el profesor ignora muchas veces la mayoría de los procesos internos, aunque hasta cierto punto sea capaz de predecir el comportamiento de su alumno. Esto es tanto más aplicable a la formación ulterior de una máquina que tenga por origen una máquina infantil con un diseño (o programa) perfectamente experimentado. Situación muy distinta al procedimiento normal de emplear una máquina para hacer cálculos, ya que el objeto, en este caso, consiste en disponer de una imagen mental clara del estado de la máquina en cada momento de la computación. Este propósito sólo es alcanzable con una imposición. La opinión de que «la máquina sólo hace lo que queramos que haga»^[3] parece extraña a la vista de lo expuesto. La mayoría de los programas que podemos introducir en la máquina la hará hacer algo que no entendemos o que consideramos como comportamiento totalmente aleatorio. El comportamiento inteligente consiste probablemente en una desviación del comportamiento absolutamente disciplinado que implica la computación, aunque relativamente leve y sin que provoque un comportamiento aleatorio o *loops* repetitivos inútiles. Otro importante resultado de la preparación de una máquina para que intervenga en el juego de imitación, merced a un proceso de enseñanza y aprendizaje, radica en que la «falibilidad humana» suele quedar descartada de una forma bastante natural, sin necesidad de «entrenamiento» especial. Los procesos que se aprenden no procuran una certeza absoluta de resultados; si así fuera, nunca fallaría su aprendizaje.

Quizá convenga introducir un elemento aleatorio en la máquina que aprende (cf. pág. 18). Un elemento aleatorio resulta bastante útil en la búsqueda de la solución de un problema. Supongamos, por ejemplo, que deseamos hallar un número entre 50 y 200 que sea igual al cuadrado de la suma de sus cifras; empecemos por el 51 y sigamos con el 52 hasta encontrar la combinación justa. Otra alternativa sería elegir números al azar hasta hallar uno que nos sirva. Este método presenta la ventaja de

que nos ahorra la necesidad de mantener el registro de los valores que se han probado, y el inconveniente de que se corre el riesgo de probar dos veces el mismo número, pero esto no es tan importante si hay varias soluciones. El método sistemático presenta el inconveniente de que puede haber una serie enorme sin solución en la región que hay que investigar en primer lugar. El proceso de aprendizaje puede considerarse como la búsqueda de una forma de comportamiento que satisfaga al profesor (o cualquier otro requisito). Como probablemente existe un gran número de soluciones satisfactorias, el método aleatorio parece mejor que el sistemático. Se advertirá que es el que interviene en el proceso análogo de la evolución, y que en ella no es posible el método sistemático. ¿Cómo sería posible conservar el registro de las distintas combinaciones genéticas ensayadas para evitar probarlas de nuevo?

Esperemos que las máquinas lleguen a competir con el hombre en todos los campos puramente intelectuales. ¿Pero cuáles son los mejores para empezar? También es una ardua decisión. Muchos piensan que lo mejor es una actividad de naturaleza tan abstracta como jugar al ajedrez. También puede sostenerse que lo óptimo sería dotar a la máquina de los mejores órganos sensoriales posibles y luego enseñarla a entender y a hablar inglés. Es un proceso que podría hacerse con arreglo al aprendizaje normal de un niño: se señalan los objetos, se los nombra, etc. Vuelvo a insistir en que ignoro la respuesta adecuada; creo que hay que experimentar los dos enfoques.

Sólo podemos prever el futuro inmediato, pero de lo que no cabe duda es de que hay mucho por hacer.

El concepto mecánico de mente^[*]

Michael Scriven

¿Existe alguna diferencia fundamental entre un hombre y una máquina? Nos han sugerido múltiples respuestas. Una de ellas consiste en afirmar que el hombre posee cualidades psicológicas, como inteligencia, conciencia u originalidad, cualidades necesariamente ausentes en la máquina. Otros ejemplos son los de introspección, pensamiento, libre albedrío, humor, amor, correlación verbal y sentidos. En este trabajo optaremos por referirnos únicamente a la conciencia, aunque la argumentación discurra de modo similar en los demás casos. Suele normalmente entenderse que una máquina es un artefacto, una invención mecánica manufacturada (y posiblemente eléctrica). Así lo entenderemos en nuestro planteamiento, siendo el propósito de la argumentación analizar en detalle la afirmación de que las máquinas nunca son conscientes.

Al decir que es imposible que una máquina sea consciente, no siempre queda claro hasta qué extremo se pretende plantear una objeción lógica, ni hasta qué extremo es empírica. Tal vez se piense que hay algún obstáculo práctico insuperable, cual es, por ejemplo, el de dotar a una máquina de medios de comunicación. Pero no hay que olvidar que es *concebible* la superación de estos obstáculos empíricos, y que un argumento basado únicamente en estas diferencias no resiste una demostración lógica. En esta discusión consideraremos que *ninguna* de esas diferencias prácticas son diferencias fundamentales, y, por lo tanto, las máquinas equiparables al hombre tendrán facultad de hablar, gesticular, deambular, etc. y serán sensibles a la luz, al sonido y a otros factores ambientales. Las denominaremos robots. Un robot puede realizar lo que cualquier máquina. ¿Hay alguna diferencia fundamental entre los robots y los seres humanos?

Se pensará que la pregunta original ha quedado desbordada por las notorias presunciones del último párrafo. De hecho, sólo es así si la cuestión de si una máquina puede ser consciente depende de su correcta emulación del comportamiento de, pongamos por caso, un hombre acerca del cual no quepan dudas de que es consciente. Un robot, *ex hypothesi*, es capaz de remedar cualquier comportamiento humano. Si un hombre responde con un respingo cuando susurramos su nombre a su espalda (o hace crítica poética, o confiesa que está enamorado), no dudamos de que sea consciente. Pero, si una máquina respondiera del mismo modo a las mismas frecuencias de sonido, nada nos convencería de que fuera consciente. La pregunta

que intentamos contestar con los tests habituales es si una persona es consciente o inconsciente; por el contrario, cuando se trata de máquinas, lo que nos planteamos es si son capaces o no de conciencia. Sabemos que la pregunta acerca de la conciencia es pertinente en el caso de una persona: lo que nos importa en el caso de una máquina no es la pregunta en sí, sino si la pregunta puede realmente plantearse. Por lo tanto, no es muy apropiado demostrar que una máquina es capaz de responder a unos tests de conciencia en el caso de una persona, sin habernos planteado si una máquina es fundamentalmente distinta a una persona en los aspectos que *no se refieren al comportamiento*, en tanto éstos sean importantes en la cuestión de la conciencia. El robot es simplemente una máquina indiferenciable del hombre en lo que al comportamiento se refiere. Por lo tanto, la pregunta clave es: «¿Es un robot fundamentalmente distinto a una persona?», y las respuestas a considerar son del tipo: «Las personas, y no los robots, son conscientes».

¿Es acaso un prejuicio intransigente lo que nos impide ampliar el concepto de conciencia a los robots, visto que éstos calculan más deprisa, reaccionan más rápido, ven más claro y recuerdan con mayor exactitud que las personas? ¿De qué carecen si lo pueden hacer todo? Hacen lo que los seres humanos, pero no pueden ser lo que son las personas: el pulso eléctrico que los activa no es la vida. Como cualquier ser encantado de las fábulas, poseen una apariencia de vida, pero no están vivos. Pasan los años y persiste nuestra decepción, pero, si por fin uno de ellos se cortase y no sangrara, o en un ataque de cólera destruyera una casa, retrocederíamos espantados exclamando: «¡Un robot, y pensar que habíamos creído que vivía y respiraba!» Nunca sintió, ni sufrió, ni pensó, ni soñó y, sin embargo, siempre dio muestras de ello. Podemos construir máquinas para arar y cosechar que ejecuten e imiten, pero no que sientan placer o autocompasión. Esto sólo son cosas posibles para los seres conscientes y, por muy ingenioso que sea el mecanismo, por muy complejo que sea el comportamiento de una máquina, aunque hable o juegue al ajedrez, nunca superará en conciencia a un reloj. Es lo que suponíamos.

Distinguiremos el significado de «consciente» por contraste con «incapaz de ser consciente» del de «inconsciente» por medio de una letra mayúscula. Con estos términos queda resuelta la aparente paradoja de los seres durmientes; ya que la inconsciencia es bastante distinto de lo Inconsciente; de hecho, lo inconsciente (y lo consciente) supone Conciencia^[1]. La pregunta «¿Es consciente?» puede plantearse para cualquier cosa. Es absurdo preguntarse de una piedra o de un cronómetro si «¿es[tá]^[*] consciente?», porque es absurdo decir que está muerto, dormido, drogado o aturdido, es decir, inconsciente. Hay casos en los que es muy difícil dilucidar la cuestión de la Conciencia. Considérense los seres vivos del mundo. Los helechos y los frutales no son Conscientes; ni las algas, ni las plantas insectívoras. Los protozoos y las bacterias no son Conscientes. Pero ¿son Conscientes las medusas y las lombrices de tierra? ¿En qué fase del desarrollo pasa el embrión humano de Inconsciente a Consciente? Se dan casos de seres vivos que no son Conscientes y hay

casos dudosos. Pero sólo en el caso de seres vivos surgen dudas. Las máquinas aparentemente no tienen vida. Incluso hay casos en los que es difícil pronunciarse sobre la cuestión de la conciencia. Por ejemplo, una persona puede tener un córtex totalmente anestesiado mientras su comportamiento lo controla un cirujano mediante ondas de radio que accionan minúsculos dispositivos implantados en los terminales nerviosos eferentes. La información de los nervios aferentes llega hasta el cirujano por medio de estos dispositivos. Esta persona está inconsciente, es un muñeco, aunque externamente sea consciente. Pero estas dificultades sólo se plantean en seres Conscientes.

Se argüirá que las máquinas se aproximan a la Conciencia por una vía distinta a la de los seres vivos: la del comportamiento. Son capaces de mayor complejidad de comportamiento que las simples criaturas que hemos citado. Se trata de otra complejidad de comportamiento (o grado de organización) vinculado fundamentalmente a la Conciencia: ¿incluso un comportamiento tan complicado como el de una persona? Se nos plantea otra vez la pregunta, pero ahora aún más difusa que antes.

Resumamos de un modo sencillo el problema. Todos sabemos lo que significa «consciente»; todos sabemos que somos y estamos conscientes cuando pensamos, recordamos, miramos o leemos. Pero no puede existir test interno alguno de las condiciones mentales de *otra* persona, y tenemos que juzgarlas externamente, *si es que podemos juzgarlas*. Aquí empiezan las dificultades. Primero, aunque podamos tener la certeza *fehaciente* de que una persona es y está consciente o inconsciente, no siempre podemos tenerla gracias al simple hecho de observarla o verificar su comportamiento: puede hallarse totalmente paralizada y sin capacidad motriz, aunque siga consciente, o puede deambular y hablar por control remoto con el cerebro anestesiado y en estado de inconsciencia. Por lo tanto, los signos externos (incluido el habla) no son signos irrefutables de conciencia y, por consiguiente, no cabe duda de que no son, incluso en personas que sabemos que son conscientes, lo mismo que conciencia. Segundo, está la cuestión, de muy distinta índole, de determinar si algo es Consciente, es decir, si es capaz de ser consciente. Sugerir que la evidencia de la conducta basta para demostrar que hay Conciencia es como sugerir que el hecho de que un ascensor empiece a subir cuando apretamos el botón demuestra que dentro hay una persona. En cada caso la evidencia es válida *tan sólo* si contamos con otra evidencia fundamental (la capacidad de inducción). Por sí sola de nada sirve, y dista mucho de ser concluyente, como tampoco se deduce a partir de un abrelatas que haya comida. La conducta no está vinculada a la Conciencia como el trueno al rayo, el sonido al eco o las nubes a la lluvia; sólo como el dolor a la tortura. Una cosa no garantiza la otra, sino que está garantizada por ella. Una serie interminable de observaciones sobre la conducta no equivale a observar la conciencia, pero puede ser un error dudar de ello como una prueba, como en el caso, por ejemplo, de una persona que conduce un coche. Una serie de observaciones sobre la conducta no

equivale a observar la consciencia, y por ello cabe dudar de que sea *siquiera* una prueba de Conciencia. Hay una relación esencial entre la capacidad de tener una conducta compleja y la Conciencia; una es condición necesaria de la otra, pero no una condición suficiente, y aunque determinemos qué seres vivos son Conscientes a partir de su conducta, no podemos establecer que cualquier cosa sea Consciente a partir de su conducta. La propia vida es condición necesaria de la Conciencia y, aunque la conducta sea un factor a veces decisivo para establecer si determinado sistema está vivo, repito que no es el único. La conducta de un sistema de predicción de artillería antiaérea es más complejo, organizado y adaptativo que el de las algas marinas, pero no por ello es algo vivo. Las algas son bioquímicamente más complejas, pero no en su modo de comportamiento. También los robots son máquinas, pero están formados tan sólo por elementos mecánicos y eléctricos, y no pueden estar vivos. Si se anunciase que los científicos han descubierto un medio para crear vida, nadie se lo tomaría en serio a menos que fueran capaces de crear una máquina compleja que lo demostrara. Es como si anunciaran que, a veces, pueden verse objetos por debajo de la línea del horizonte y lo demostraran con la definición científica que utilizan astrónomos y observadores, según la cual cualquier objeto a un nivel más bajo que la línea de visión horizontal está por debajo del horizonte: el suelo que pisamos, por ejemplo. El hecho de que, en las máquinas y en los organismos vivos, se dé la misma característica de unidad orgánica rectora no basta para probar que ambos viven. No es tan sorprendente que seamos capaces de construir un mecanismo que posea esta propiedad relativamente simple, mientras que sería extremadamente sorprendente que el producir semejante mecanismo le confiriera vida. Otros tests notables también fallan por simples que sean. Al igual que es absurdo hablar de una máquina muerta, también es absurdo decir que está viva. Las máquinas ni siquiera pertenecen a la *categoría* de cosas muertas o vivas. Por supuesto podría utilizarse materia orgánica para construir un robot, pero esto no modificaría un ápice la argumentación. Aunque se demostrase la posibilidad de crear células vivas y de integrarlas al artillugio, el resultado no sería una máquina consciente, ya que una proporción suficiente como para justificar en el objeto (y no partes de él) el calificativo de viviente y, por lo tanto, posiblemente de consciente, bastaría sin duda para descartar por incorrecta la denominación de máquina. Este es el último razonamiento y, como los anteriores, no nos autoriza a eludir la conclusión de que no puede existir nada que pueda calificarse de máquina consciente.

Conclusión

Parece darse una paradoja asociada al concepto de máquina consciente. Por un lado, no parece que haya nada en la construcción, en los elementos o en la conducta del ser humano que sea rotundamente imposible de duplicar o sintetizar científicamente. Por otro, parecen existir ciertas descripciones fundamentales y

significativas de conducta humana que no pueden en rigor aplicarse a las máquinas. Nos causa perplejidad el que la base para una descripción pueda reproducirse, aunque no pueda volver a aplicarse la descripción.

Este conflicto hizo que en su momento se negase la composición material del cuerpo y, en épocas más recientes, que se niegue que el ser humano posee atributos que una máquina jamás podrá igualar.

Pero sólo los rompecabezas y los problemas tienen solución: las paradojas, no; si acaso, a veces, resoluciones. Cuando se resuelve una paradoja no se pronuncia uno por ninguno de los argumentos contradictorios, sino que se extrae la auténtica verdad de cada uno de ellos para demostrar que su contradicción es de distinto nivel. Esto es aplicable a las paradojas habituales, lógicas y pragmáticas, así como a las paradojas que constituyen la base de problemas filosóficos como el que estamos exponiendo. La argumentación nos revela que el ser humano no encierra necesariamente un elemento transcendente, y, sin embargo, las máquinas nunca serán conscientes, porque hemos visto que la reproducción de un hombre lo suficientemente exacta como para ser consciente sigue siendo una máquina, por exacta que sea.

Aplicaciones

Hay que discutir aparte la aplicación de los resultados analíticos como los aquí expuestos. La cuestión a dilucidar es si cualquier posible máquina es esencialmente distinta a una persona. Una de las posibles respuestas nos sugiere que una máquina nunca será consciente. Las palabras clave son «máquina» y «consciente», pero es muy posible que cambie el significado de estas palabras y, en consecuencia, que sus definiciones no guarden ya la misma relación lógica con otras, como actualmente es el caso. Podría suceder esto por diversos motivos, y el más viable sería que las asociaciones ineludibles que provocan indujeran a emplearlas en contextos en los que su sentido estricto y una asociación menos forzada las hicieran inadecuadas o incorrectas. Tenemos un ejemplo de ello en el caso de la palabra «inteligente». Es perfectamente imaginable una persona cuyo trabajo consista fundamentalmente en manejar una de las más potentes computadoras electrónicas, a la cual aplica tal calificativo. La persona comete frecuentes errores, la máquina es exacta. La persona tiene una limitación memorística de cifras; la máquina una enorme capacidad de almacenamiento. La persona es inteligente, pero la máquina trabaja mejor. Al principio, a guisa de jerga, y después en serio, estas máquinas serán calificadas de inteligentes. Quizá lleguemos a idear el modo de tasar la inteligencia de las distintas máquinas más por su velocidad y su precisión que por su mera capacidad, o quizá también por su versatilidad y su simplicidad de programación. Conforme se construyan mejores computadoras y las vayamos empleando más para resolver problemas íntegros en vez de para realizar cálculos específicos, se irá imponiendo el concepto de *consultar* la computadora sobre el de *usarla*. En otros muchos campos su

utilización reflejará la creciente tendencia a considerar la computadora como al especialista *par excellence*. Entonces, alguien se planteará la pregunta: «¿Llegarán las máquinas a ser realmente inteligentes?» Pero ya no es igual que los interrogantes que planteábamos al principio, aunque las palabras sean las mismas. No hemos dado una respuesta; ni siquiera un esbozo: es factible considerar todas las máquinas posibles y absurdo considerar todos los posibles significados de las palabras «máquina» o «inteligencia».

Sin embargo, hay circunstancias en las que es difícil determinar si las descripciones utilizadas o las afirmaciones que se hacen encierran un cambio de sentido en sus términos o simplemente representan una combinación de términos sin precedente, pero no por ello impropia. Un ejemplo de combinación sin precedente, que no implicara cambio de sentido, sería la descripción «uva azul». Acabamos de exponer un ejemplo de cambio de significado al sugerir que la palabra «inteligente» modificara su sentido al aplicarse a una computadora. Según su actual significado, la palabra inteligente no sirve para calificar una regla de cálculo o un abaco, y el hecho de que los accione un motor en lugar de las manos de una persona no contribuye a recibir la categoría de inteligentes. Una computadora electrónica es un complicado instrumento del que nos servimos con el mismo propósito que de una regla de cálculo o de un ábaco. A menos que la accione un ser inteligente, es totalmente inútil; pero no porque la accione un ser inteligente se convierte en inteligente, tan sólo se hace útil. Su construcción puede ser totalmente similar a la del cerebro humano, pero el cerebro humano es un mecanismo fisiológico, no más inteligente que un músculo. No obstante, desde luego, el cerebro, a diferencia del músculo, es un elemento indispensable al ser inteligente, y la relativa inteligencia de los distintos seres se deduce a partir de su respectiva estructura cerebral; pero es el ser el que es inteligente, no su cerebro. Las personas tienen buenos cerebros o cerebros mediocres, pero nunca cerebros inteligentes. Ahora bien, las computadoras no ocupan siquiera un lugar tan privilegiado como los cerebros en esta competición, ya que no son indispensables para la inteligencia de nadie, del mismo modo que los libros no son indispensables para la memoria. Y nadie puede decir que una computadora es inteligente ni que un libro se acuerda de cosas. Este es el argumento que demuestra que «inteligente» es un adjetivo inapropiado para el sustantivo «computadora». También es un argumento que demuestra que se ha producido un cambio de significado en uno de estos dos términos, si la combinación se usa con propiedad. En la situación descrita anteriormente, se da un uso apropiado, aunque quizás estrictamente técnico, de la descripción. Por lo tanto, ha habido un cambio de significado. De igual modo, si consideramos un ejemplo simple, podemos decir con bastante certeza que, si existe un empleo estándar de la expresión «padre (o madre) sin hijos», es que ha cambiado el significado de uno de los términos.

En la primera parte de este ensayo se argüía que la expresión «máquina consciente» no podía tener un sentido apropiado; sugeríamos que, en el empleo

normal, los factores que contribuyen a la identificación de las máquinas, son también factores que las diferencian de la materia animada y, *a fortiori*, de los seres conscientes. A la inversa, los factores que nos inducen a emplear el adjetivo «consciente» incluyen, explícita o implícitamente, el requisito de que el sujeto descrito no sea un artilugio mecánico. Si en este punto nos dejamos inducir a error, el consiguiente descubrimiento de su estructura inorgánica y de su accionamiento eléctrico nos obliga a descartar el calificativo, del mismo modo que descartaríamos el adjetivo «enfadado» en la descripción de una acción realizada por una persona que descubriéramos se limita a fingir. No existe diferencia entre el comportamiento de un robot y el de una persona consciente, ni entre el de un actor y el de un hombre enfadado, pero el robot no es realmente consciente, del mismo modo que el actor no está realmente enfadado. De ello se deduce que, si empleamos la expresión de «máquina consciente», uno de los términos ha cambiado de significado.

Pero se da la situación inversa en la que en principio no es que nos falte la información adecuada, pero sí que nos encontremos de pronto con el hecho consumado, con lo que era imposible ya realizado. Unos exploradores espaciales aterrizan en otro planeta y regresan a nuestro mundo con un habitante extraterrestre: un robot. Este forma parte de una civilización próspera y desarrollada; es un ciudadano hábil y culto, de aspecto antropeide, pero con poderes poco corrientes. Pasa varios meses en la Tierra observando con interés nuestros sistemas sociales y económicos. En tales circunstancias cabe esperar que muchas personas que tengan la oportunidad de conocerle, hablen de él como si se tratara de una persona. Supongamos que alguien lanza la idea de que el robot es un ejemplar de máquina consciente. No es tan sencillo detectar en tal descripción una incongruencia fundamental como en la citada anteriormente sobre la aplicación de la expresión «padre (o madre) sin hijos», o incluso en la de «computadora inteligente».

Pero consideremos ahora cómo reaccionaría el escéptico irreductible que se negara a aceptar que el robot es una máquina consciente. Podemos señalar que el robot ha permanecido un tiempo en la Tierra y se ha comportado como un amigo y un colega culto. Nos preguntamos cómo podría ser posible si el robot no fuera un ser consciente, un ser auténticamente capaz de sentir simpatía, reflexionar y poseer razonamiento interiorizado. En tal caso no hay posibilidad de que el escéptico repita la técnica expositiva de la imposibilidad de inteligencia en una computadora, ya que el robot es un ser autónomo y no un instrumento en sentido normal; y a algunos seres autónomos se les puede con propiedad denominar «conscientes».

El escéptico se plantea en primer lugar si se le puede realmente calificar de máquina. Ciertamente —se dice—, es metálico y de aspecto mecánico, por fuera y por dentro. Quizás estas solas características definan una máquina, pero desde luego no basta con que tenga la apariencia. Hay que estar seguros de que a todos los niveles el comportamiento de sus componentes y, en gran medida, del robot, es mecánico. Además, puede que el origen real de sus movimientos, de su habla y de sus gestos

consista en alguna fuente no mecánica, quizás un alma o un ser extraño. Además —se dice—, ni aun eso bastaría para entender la naturaleza general de cada conexión: para asegurarse de que este robot es tan sólo una máquina, hay que comprobar que todos sus elementos son cuantitativamente mecánicos y que la magnitud de cada reacción es perfectamente deducible del parámetro de cada estímulo relacionado con la naturaleza de las conexiones. De hecho, habría que ver claramente que no es más que una simple máquina, que no posee «voluntad» propia. Si lo hubiéramos construido nosotros, la mayoría de la gente quedaría, naturalmente, satisfecha con este aspecto del planteamiento. Supongamos que logramos analizar correctamente el mecanismo, en cuyo caso podríamos prever fácilmente todos los movimientos del robot. Lo más probable, como sucede con las grandes computadoras, es que, aunque *en principio* fuera posible predecir todos los movimientos, rara vez podría hacerse con la misma rapidez que la máquina. En cierto sentido puede decirse que el comportamiento de los seres humanos es predecible, si nos referimos a la convicción que tenemos de que no son una excepción al concepto materialista de la naturaleza, con lo cual no se sugiere que, con los actuales conocimientos, se pueda predecir el comportamiento. En un sentido ligeramente distinto, podríamos decir que el comportamiento de una ruleta o la súbita avería de un motor son, en principio, predecibles. Cuando decimos que no podemos asegurar con certeza que el robot es una máquina si no estamos seguros de que su comportamiento es en principio predecible, queremos decir que, con el tiempo, podríamos calcular el comportamiento que resulte de cualquier estímulo ambiental. Esto no significa que podamos en todo momento decir lo que hará el robot en una determinada situación antes de que lo haga.

Ahora bien, la relación entre Conciencia y libre albedrío es bastante complicada. No es fácil demostrar que una sea imposible sin el otro, pero, cuando menos, es seguro que no podemos demostrar que el robot es Consciente, ya que disponemos de una explicación totalmente causal de toda su gama de comportamientos; esta explicación en modo alguno depende de su conciencia, por lo que no puede darse su comportamiento como prueba de que posee conciencia. La Conciencia no es una propiedad que se detecte en una máquina por medio de un examen físico, porque no puede *identificarse* con ninguna de las características físicas de una máquina. Tampoco puede correlacionarse con ninguna de esas características, como se correlaciona el color rojo con determinadas longitudes de onda luminosa. El factor fundamental en esta correlación es la impresión subjetiva, y precisamente lo que se cuestiona es la presencia en el robot de este elemento. Si, por ejemplo, advirtiéramos en el robot aumentos del ritmo alfa, que casi siempre corresponden al inicio de una actividad verbal, cabría caer en la tentación de afirmar que es un individuo con conciencia. Sin embargo, también podría calificarse de signo preparatorio mecánico para el habla, u otra actividad.

Sólo los filósofos han puesto en duda el que se pueda o no demostrar que otras personas son conscientes. Esta postura se debe, en parte, a las fuentes de error que

conlleva el diagnóstico de ciertos estados mentales del comportamiento de otros individuos, ausentes en el conocimiento de estos estados mentales en uno mismo. Es una duda filosófica, no científica, porque no hay (ni puede haber) correlación entre la mencionada diferencia y ninguna diferencia científicamente observable. La duda de si un robot puede ser consciente no es sólo filosófica, sino científica, porque hay diferencias físicas (no de comportamiento) entre el hombre y el robot.

En realidad es un caso bastante más flagrante, diría el escéptico, puesto que es curiosamente endeble la afirmación de que un robot puede ser consciente aunque su comportamiento esté totalmente determinado (tenga o no libre albedrío). Supongamos que el robot sugiriera que la cuestión de su conciencia la sabe él mejor que nadie, diera una buena explicación de lo que significa la palabra y afirmara que es consciente. Como conocemos perfectamente su mecanismo, podría pensarse que es posible determinar si miente con el fin de decidir en definitiva la cuestión de su conciencia. Pero, para decidir si está o no mintiendo al afirmar que es consciente, hay que decidir primero si es consciente o no, es decir, contestar a la pregunta original. Para decidir si los seres humanos sienten cuando nosotros ignoramos la verdad, podemos emplear un detector de mentiras, pero éste depende de los efectos subsiguientes a la afirmación consciente de una falsedad, y en el robot resultaría imposible esta clase de prueba sin antes haber concretado si es o no consciente. Un gramófono puede decir mentiras, pero no miente, aunque tampoco dice la verdad. Por lo tanto, el testimonio del robot no tiene validez, porque no podemos demostrar que es su propio testimonio. Un robot, a diferencia de un gramófono, tiene voz propia, pero lo que dice depende de su construcción y aplicación. Vuelve a asaltarme la duda de si es admisible hablar de máquina cuando no podemos afirmar si fue diseñado, y para qué, si posee antepasados y descendientes, si fue concebido por diseñadores o por mecánicos.

En la primera parte de este trabajo argumentamos en base a una idea bastante simple (un concepto no sometido a análisis) de máquina y organismo vivo. Consideramos asimismo el caso de un robot recién descubierto, con antecedentes desconocidos, en el que deseábamos encontrar las condiciones mínimas para denominarlo máquina y decir que está Consciente. Pese a decidir que es una máquina en sus más mínimos detalles, vemos que todavía parece posible *decir* que el robot es Consciente, pero entonces la Conciencia se ha convertido en algo totalmente distinto al comportamiento, un Observador fantasmagórico de cuya presencia no hay indicios y que no influye para nada. Es falso decir que la conciencia es *necesaria* para explicar determinado comportamiento en las máquinas (por ejemplo, el empleo de determinado lenguaje), aunque pueda parecer que la conciencia *va a la par* con determinado comportamiento de las máquinas y no *se deduzca* de él, como sucede con las personas. Las máquinas, que parecen emplear correctamente la palabra «consciente», lo hacen sencillamente porque están programadas de un modo específico: si hubiera otra explicación no serían simples máquinas.

En este caso los antecedentes de las características desconocidas cobran importancia. Si viéramos a los robots en su planeta de origen, creciendo en la niñez, aprendiendo a pensar y a hablar, convirtiéndose de algún modo en padres y muriendo, podríamos convencernos de que las categorías habituales de «máquina» y «organismo vivo» requieren quizás el complemento de una tercera, la de «androide». En tal caso sería señal de un grave desconocimiento de la situación de los nuevos conceptos preguntar si los androides son realmente máquinas, aunque podamos calificarlos de Conscientes. Si, en cambio, salieran prefabricados de fábricas automatizadas y, como cualquier producto prefabricado, fueran exterior-mente idénticos, con la misma expresión facial, las mismas facultades y el mismo vocabulario, quizás consideraríamos que es correcto llamarlos «máquinas». Quizá dijésemos también, incluso en este caso, que son Conscientes, ya que el hombre tiende muchas veces a personificar las máquinas con las que está familiarizado: pero no sería realmente una razón de peso para creer que son auténticamente Conscientes. Si no dispusiéramos de una explicación causal de su comportamiento, si no hubiera en su interior válvulas y ruedas, y sólo viésemos un plástico homogéneo o una cavidad, la calificación de Conciencia sólo se comprendería como un recurso ignorante y senil, un tic antropocentrista. Tal como son, su comportamiento es prueba de su construcción, no de su Conciencia. Ahora bien, es cierto que nuestro conocimiento del mecanismo robótico no hace que su comportamiento sea más previsible, ni tampoco más humano (ya que hemos asumido que actúan como seres humanos), por lo que sentimos la tentación de justificar lo imprevisible y lo humano afirmando que son Conscientes. Sin embargo, parecen humanos por la sencilla razón de que fueron contruidos para que parecieran humanos. Si existe algún motivo para dudar de que hayan sido diseñados y contruidos, como en el caso de los androides, los consideramos conscientes, pero no máquinas. Si no existe ese motivo, los consideramos máquinas, pero atribuirles Conciencia es introducir la sombra de un fantasma en la máquina.

Addendum (1963)

Considérese este trabajo un testimonio (en parte redactado como reacción) de las dificultades que entraña la asimilación que hace Turing del conductismo al campo de las computadoras [cf. el primer artículo en este volumen]. Desde la fecha en que fue escrito, he podido comprobar que pueden solventarse muchas de estas dificultades de muchas maneras. Se puede, por ejemplo, obtener un apoyo inductivo abrumador con respecto a la veracidad de un agente (artificial o no), a partir de campos que no afecten sus manifestaciones introspectivas, extrapolándola en apoyo de una manifestación de conciencia del dolor, a condición de proceder con sumo cuidado al enseñar el lenguaje de las sensaciones [cf. «The Compleat Robot: A Preface to Androidology», citado en la bibliografía]. Además, actualmente parece posible

definir los elementos de que debe constar un sistema computerizado capaz de una auténtica comprensión del lenguaje [cf. *Computers and Comprehension*, publicaciones RAND], Creo que actualmente se puede construir una supercomputadora que niegue de un modo totalmente razonable la afirmación de que no tienen sentimientos.

Mentes, máquinas y Gödel^[*]

J. R. Lucas

En mi opinión, el teorema de Gödel demuestra la falsedad del Mecanicismo, o sea, que las mentes no pueden ser descritas como máquinas. Esta es también la opinión de otros muchos, y prácticamente todos los matemáticos a quienes les he planteado la cuestión han reconocido sostener una postura parecida, aunque se hayan negado a comprometerse rotundamente desoyendo argumentos que planteen y razonen todas las objeciones^[1]. Lo intentaré en este trabajo.

El teorema de Gödel enuncia que, en cualquier sistema coherente de suficiente potencia para producir simple aritmética, hay fórmulas que no pueden demostrarse dentro del sistema, pero que pueden aceptarse como ciertas. Consideremos en particular la fórmula que dice: «Esta fórmula es indemostrable dentro-del-sistema». Si esta fórmula fuera demostrable-dentro-del-sistema, tendríamos una contradicción, ya que, si fuera demostrable-dentro-del-sistema, no sería indemostrable-dentro-del-sistema, de modo que «esta fórmula es indemostrable-dentro-del-sistema» sería falsa y, también, si fuera demostrable-dentro-del-sistema, no sería falsa sino verdadera, ya que, en cualquier sistema coherente, nada falso puede demostrarse-dentro-del-sistema, salvo las verdades. Por lo tanto, la fórmula: «Esta fórmula es indemostrable-dentro-del-sistema» no es demostrable dentro del sistema, sino indemostrable-dentro-del-sistema. Además, si la fórmula: «Esta fórmula es indemostrable-dentro-del-sistema» es indemostrable-dentro-del-sistema, entonces es cierto que la fórmula es indemostrable-dentro-del-sistema, es decir, que «Esta fórmula es indemostrable-dentro-del-sistema» es verdad.

La argumentación expuesta es muy enrevesada y difícil de entender del todo. Es preferible invertir el argumento: considerar la posibilidad de que «Esta fórmula es indemostrable-dentro-del-sistema» puede ser falsa, demostrar que es imposible y con ello que la fórmula es verdad; de lo que se sigue que es indemostrable. Aun así, el argumento sigue siendo muy poco convincente: da la sensación de que encierra alguna trampa. La única función del teorema de Gödel es demostrar que no hay trampa alguna y que el resultado puede obtenerse por la más rigurosa deducción; es válido para todos los sistemas formales que sean (a) coherentes y (b) adecuados para la aritmética simple, es decir, que contengan los números naturales y las operaciones de adición y multiplicación. O sea que contengan fórmulas indemostrables, aunque

perfectamente significativas, algunas de las cuales, además, nosotros, que estamos fuera del sistema, vemos que son verdad.

El teorema de Gödel debe ser aplicable a las máquinas cibernéticas, porque es fundamental para la condición de máquina el ser un ejemplo concreto de un sistema formal. De ello se deduce que, dada una máquina que sea coherente y capaz de efectuar operaciones aritméticas simples, existe una fórmula cuya autenticidad es incapaz de demostrar (o sea la fórmula es-indemostrable-dentro-del-sistema), aunque nosotros veamos que es cierta. De ello se infiere que ninguna máquina puede ser un modelo exacto o adecuado de la mente, y que las mentes son fundamentalmente distintas a las máquinas.

Entendemos por máquina cibernética un aparato que efectúa una serie de operaciones con arreglo a una serie concreta de reglas. Normalmente «programamos» la máquina, es decir, le damos una serie de instrucciones relativas a lo que debe hacer en cualquier eventualidad, y la alimentamos con la «información» inicial por medio de la cual efectúa los cálculos. Al considerar la posibilidad de que la mente sea un mecanismo cibernético, nos imaginamos el mismo esquema; suponemos que el cerebro está formado por circuitos neurológicos complejos y que la información que aportan los sentidos «se procesa» y elabora o se almacena para el futuro. Si es éste el mecanismo, entonces, dados el modo en que se programa —el modo en que está «instalado»— y la información que recibe, la respuesta —la «salida»— está determinada y podría perfectamente calcularse si dispusiéramos de suficiente tiempo. Nuestro concepto de máquina se reduce a algo cuyo comportamiento está totalmente determinado por su modalidad constructiva y por los «estímulos» que recibe, sin que haya posibilidad de que actúe por su cuenta: dadas una modalidad constructiva y una determinada entrada de información, tiene que actuar de un modo concreto. Sin embargo, no deberíamos interesarnos por lo que *debe* hacer una máquina, sino por lo que *puede* hacer. Es decir que, en lugar de considerar el conjunto de reglas que determinan con exactitud lo que la máquina hace en determinadas circunstancias, debemos considerar únicamente un esbozo de las reglas que delimiten las posibles respuestas de la máquina, pero no del todo. Las reglas completas determinarán del todo las operaciones en cada fase; en cada fase habrá una instrucción concreta, por ejemplo: «Si el número es primo y mayor de dos, suma uno y divide por dos; si no es primo, divídelo por su factor más pequeño». Sin embargo, consideremos la posibilidad de que existan instrucciones alternativas, por ejemplo: «En un quebrado, divide numerador y denominador por *cualquier* número que sea factor de ambos». Flexibilizando así la especificación del modelo, de manera que no sea totalmente determinista, aunque siga siendo totalmente mecánico, podemos incorporar una característica propuesta muchas veces en modelos mecánicos de la mente, o sea que contenga un atisbo de aleatoriedad. Se puede construir una máquina en la que la opción entre una serie de alternativas esté determinada por, pongamos por caso, el número de átomos de radio que se hayan desintegrado en un depósito concreto

durante el último medio minuto. Es plausible *prima facie* que nuestro cerebro esté sujeto a efectos aleatorios: un rayo cósmico puede perfectamente desencadenar un impulso neuronal. Pero está claro que, en una máquina, no se puede introducir un dispositivo aleatorio que opte por cualquier tipo de alternativa: sólo tiene capacidad para elegir entre el número de alternativas permitidas. Está muy bien sumar *cualquier* número elegido al azar a los términos de una ecuación, pero no es lo mismo que sumar un número a un término y otro distinto al otro. Está muy bien elegir para la demostración un teorema de Euclides en vez de otro, o seguir un método concreto, pero no es lo mismo «demostrar» algo que no es cierto, o utilizar un «método de demostración» que no sea válido. Todos los dispositivos aleatorios deben permitir opciones únicamente en el transcurso de aquellas operaciones que no produzcan incoherencia: que es exactamente lo que especifica la instrucción flexible del modelo citado. Efectivamente, podemos expresarlo de este modo: en lugar de considerar lo que una máquina totalmente determinada *debe* hacer, consideraremos lo que esta máquina es capaz de hacer si dispusiera de un dispositivo aleatorio que actuase siempre que hubiera dos o más operaciones posibles, ninguna de las cuales diera resultados incoherentes.

Si se construyera una máquina así para producir teoremas aritméticos (en muchos aspectos, la parte más simple de las matemáticas), sólo tendría un número finito de componentes y, por lo tanto, sólo podría efectuar un número finito de tipos de operación y sólo un número finito de asunciones iniciales sobre las que operar. Efectivamente, podríamos incluso decir que sólo habría, a nivel constructivo, un número *definido* de tipos de operación y de asunciones iniciales. Las máquinas son definidas, y no consideraríamos máquina a algo indefinido o infinito. Obsérvese que decimos número de *tipos* de operación, no número de operaciones. Con suficiente tiempo y a condición de que no se gastara, una máquina podría repetir cualquier operación indefinidamente; la limitación estriba en que sólo puede haber un número definido de distintas *clases* de operaciones para que la máquina actúe.

Si sólo hay un número definido de tipos de operación y asunciones iniciales estructuradas en el sistema, podemos representarlas mediante los símbolos correspondientes inscritos en papel. Podemos simultanear la operación con reglas («reglas de inferencia» o «esquemas axiomáticos») que nos permitan pasar de una o más fórmulas (o ni siquiera a partir de una fórmula) a otra fórmula, y podemos simultanear las asunciones iniciales (si las hay) con una serie de fórmulas iniciales («proposiciones», «postulados» o «axiomas originales»). Una vez representados gráficamente, podemos también representar sobre papel cualquier operación: basta con dar fórmulas que representen la situación antes y después de la operación, anotando qué regla se sigue. Por lo tanto, podemos representar sobre papel cualquier secuencia posible de operaciones que realice la máquina. Por mucho tiempo que invierta la máquina en las operaciones, podríamos, con suficiente tiempo, papel y paciencia, escribir un análogo de las operaciones de la máquina. Este análogo sería de

hecho una prueba formal en la que cada operación de la máquina estaría representada por la aplicación de una de las reglas, y en la que las condiciones, que determinan si la máquina puede efectuar una operación en determinada circunstancia, se convierten —en esa representación— en requisitos que estipulan si puede aplicarse una regla a una fórmula concreta, o sea condiciones formales de aplicabilidad. Así, estructurando las reglas como reglas de inferencia, obtendremos una secuencia probatoria de fórmulas, inscrita cada una de ellas en virtud de una regla formal de inferencia que haya sido aplicada a otra fórmula o fórmulas previas (salvo, naturalmente, las fórmulas iniciales, que se dan porque representan asunciones iniciales integradas en el sistema). Por lo tanto, las posibles conclusiones verídicas que produzca la máquina corresponderán a los teoremas susceptibles de demostración en el sistema formal correspondiente. Integremos ahora una fórmula gödeliana en este sistema formal. Esta fórmula no puede *demostrarse-dentro-del-sistema*. Por lo tanto, la máquina no puede producir la correspondiente fórmula verídica, pero *sabemos* que la fórmula gödeliana es verdad: cualquier ser racional puede seguir el razonamiento de Gödel y convencerse de que esta fórmula, aunque indemostrable-en-el-sistema-propuesto, es, no obstante —en realidad precisamente por esta razón— cierta. Ahora bien, cualquier modelo mecánico de la mente debe incluir un mecanismo capaz de enunciar verdades aritméticas, porque es algo que pueden hacer las mentes; de hecho, es fácil construir modelos mecánicos que en muchos aspectos produzcan verdades aritméticas mucho mejor que los seres humanos. Pero en este aspecto concreto no pueden hacerlo tan bien, ya que para cada máquina hay una verdad cuya veracidad no puede demostrar, mientras que una mente sí puede. Esto demuestra que una máquina no puede ser un modelo completo y adecuado de la mente. No puede hacer *todo* lo que hace una mente, ya que, por mucho que haga, siempre hay algo de lo que es incapaz y de lo que una mente es capaz. Esto no quiere decir que no podamos construir una máquina capaz de simular *cualquier* patrón de comportamiento mental, lo que sucede es que no podemos construir una máquina que simule *todo* tipo de comportamiento mental. Podemos (o podremos algún día) construir máquinas capaces de reproducir fragmentos de comportamiento mental y, por supuesto, de mejorar el rendimiento de la mente humana, pero, por muy buena que sea la máquina y por mucho que mejore en todos los aspectos la actuación de la mente humana, siempre tendrá ese punto débil: ser incapaz de hacer ese algo que la mente sí puede. La fórmula gödeliana es el talón de Aquiles de la máquina cibernética, y por ello no cabe esperar que podamos construir una máquina capaz de hacer todo lo que hace la mente: nunca podremos, ni siquiera en principio, conseguir un modelo mecánico de la mente.

Esta conclusión resultará para algunos un tanto sospechosa, y la primera objeción será que es incompatible decir que una máquina *puede* simular *cualquier* tipo de comportamiento mental y no *puede* simular todo tipo de comportamiento. Para algunos resulta contradictorio; me limitaré a señalarles que no hay contradicción

entre el hecho de que para cualquier número natural pueda hallarse un número mayor y el hecho de que no pueda darse un número mayor que todos los números.

Podemos recurrir a la misma analogía en refutación de los que, al hallar una fórmula de la que su primera máquina no puede demostrar la veracidad, concluyen que la máquina es inadecuada y acto seguido tratan de construir otra más adecuada, en la que *puede* demostrarse la veracidad de la fórmula. Efectivamente, pueden hacerlo, pero esta segunda máquina contendrá su propia fórmula gödeliana, integrada por la aplicación del procedimiento de Gödel al sistema formal que representa su propio esquema ampliado (de la segunda máquina) de operaciones. Y esta segunda máquina no podrá demostrar la veracidad de dicha fórmula, mientras que una mente es capaz de comprender que es cierta. Si construimos una tercera máquina capaz de hacer lo que era imposible para la segunda, sucederá exactamente igual: habrá una tercera fórmula, la fórmula gödeliana para el sistema formal correspondiente al esquema operacional de la tercera máquina, que ésta será incapaz de demostrar, mientras que una mente seguirá siendo capaz de comprender su veracidad. Y así sucesivamente. Por compleja que sea la estructura de una máquina, ésta siempre corresponderá, por tratarse de una máquina, a un sistema formal que a su vez estará sujeto al procedimiento de Gödel de hallar una fórmula indemostrable-en-ese-sistema. Fórmula que la máquina no podrá demostrar que es cierta, mientras que una mente sí entenderá que lo es. Así pues, la máquina seguirá sin ser un modelo adecuado de la mente. Tratamos de construir un modelo de la mente mecánico — esencialmente «muerto»—, mientras que la mente, al ser «viva», siempre superará a cualquier sistema formal, osificado y muerto. Gracias al teorema de Gödel la mente tiene siempre la última palabra.

Pasaremos a hacer una segunda objeción. El procedimiento por el que se construye la fórmula gödeliana es un procedimiento estándar, puesto que sólo así podemos estar seguros de poder integrar esta fórmula en cualquier sistema formal. Pero, si es un procedimiento estándar, debe ser posible programar la máquina para que lo cumpla. Podríamos construir una máquina que llevara las operaciones habituales, y, además, la operación que consiste en someterse al procedimiento de Gödel, que diera a continuación la conclusión de la veracidad de este procedimiento; y así sucesivamente cuantas veces fuera necesario. Esto equivaldría a contar con un sistema dotado de una regla suplementaria de inferencia que nos permitiera añadir, a guisa de teorema, la fórmula gödeliana del resto del sistema formal, y luego la fórmula gödeliana del nuevo sistema formal reforzado, y así indefinidamente. Sería equivalente a añadir al sistema formal original una secuencia infinita de axiomas que fueran cada uno de ellos la fórmula gödeliana del sistema obtenido previamente. Aun así, no queda zanjada la cuestión, puesto que la máquina con un operador gödelófilo, por así decirlo, es una máquina *distinta* a las máquinas que no tengan esa clase de operador, y, aunque la máquina con tal operador fuera capaz de hacer las cosas en las que las máquinas sin el operador quedan superadas por una mente, cabría esperar que

una mente, enfrentada a una máquina que contase con operador gödelófilo, lo tuviera en cuenta y superara gödelianamente la nueva máquina, gödelizando operador y todo. Así resultó de hecho en la práctica. Incluso si añadimos a un sistema formal la serie infinita de axiomas formada por las sucesivas fórmulas gödelianas, el sistema resultante sigue siendo incompleto y contiene una fórmula indemostrable-dentro-del-sistema, que, sin embargo, un ser racional externo al sistema sí puede demostrar^[2]. Era de esperar, ya que, aunque añadamos una serie infinita de axiomas, hay que especificarlos mediante alguna regla finita, y una mente que considere el sistema formal ampliado puede tener en cuenta esta nueva regla de especificación. En cierto modo, por el hecho de que la mente tiene la última palabra, siempre puede descubrir un lapsus en cualquier sistema formal que se le presente como modelo de su propia actividad. En cierto sentido, el modelo mecánico ha de ser finito y definido y, en consecuencia, la mente siempre puede superarlo.

Esta es la respuesta a la objeción planteada por Turing al argumentar que la limitación de los poderes de una máquina no es una objeción de gran peso. Aunque cada máquina por separado sea incapaz de dar la respuesta adecuada a ciertas preguntas, al fin y al cabo cualquier ser humano individualmente es también falible y, en cualquier caso, «nuestra superioridad sólo es evidente en tal caso en relación con la máquina respecto a la cual nos hemos apuntado nuestro modesto triunfo. No se puede vencer a *todas* las máquinas a la vez». Pero no se trata de eso. No estamos discutiendo si las mentes o las máquinas son superiores, sino si son iguales. En ciertos aspectos las máquinas son sin duda superiores a las mentes humanas; está claro que el aspecto en que fallan es bastante poco importante y hasta trivial, pero suficiente para demostrar que una máquina *no es igual* a una mente. Ciertamente que la máquina puede hacer muchas cosas que una mente es incapaz de hacer, pero, en último extremo, si hay algo que la máquina no puede hacer y la mente sí, por trivial que sea, no podemos equipararlas, ni cabe esperar que haya nunca un modelo mecánico que sustituya convincentemente la mente. Tampoco significa esto que se trate únicamente de una máquina concreta a la que hemos vencido, ya que este triunfo no se limita a *una* máquina aislada, sino que se obtiene sobre *cualquier* ente que se especifique —en latín *quivis* o *quilibet*, no *quidam*—, y un modelo mecánico de la mente tiene que ser una máquina concreta, si bien sea cierto que cualquier «triunfo» concreto de una mente sobre una máquina puede «aventajarlo» otra máquina capaz de dar la respuesta que la primera máquina no pudo dar, de forma que «no se puede vencer a todas las máquinas a la vez», pero es algo irrelevante. Lo que está en juego no es la competición desigual entre una mente y todas las máquinas, sino el hecho de si puede haber una máquina capaz de hacer todo lo que hace la mente. Para que la tesis mecanicista se mantenga, debería ser posible, en principio, construir un modelo, uno solo, capaz de hacer todo lo que hace la mente. Es como un juego^[3]. El mecanicista inicia la partida y produce *un cualquier*, modelo mecánico de la mente, pero sólo *uno concreto*. Señalo algo que yo no puedo hacer, pero la mente sí. El

mecanicista puede modificar su modelo, pero cada vez que lo haga puedo buscar defectos en el modelo revisado. Si el mecanicista es capaz de inventar un modelo en el que yo no pueda encontrar defectos, su tesis se confirma y, si no puede, al fallar la demostración como realmente falla, queda refutada. Para triunfar tiene que ser capaz de producir un modelo mecánico definido de la mente, el que quiera, pero que pueda especificarse y que tenga la obligación de ajustarse a él. Pero, como no puede —no puede en principio— producir modelo mecánico adecuado alguno, aunque el punto débil sea insignificante, tiene que fracasar, y el mecanismo resulta necesariamente falso.

Pueden hacerse objeciones más fundadas. El teorema de Gödel es aplicable a los sistemas deductivos, y los seres humanos no están sujetos a la limitación de efectuar únicamente reflexiones deductivas. El teorema de Gödel es aplicable únicamente a sistemas coherentes, y cabe dudar de hasta qué punto se puede creer que los seres humanos son coherentes. El teorema de Gödel es únicamente aplicable a sistemas formales, y no existe *a priori* limitación al ingenio humano que descarte la posibilidad de que podamos inventar una réplica humana que no sea representable mediante un sistema formal.

Los seres humanos no están sujetos a la limitación de efectuar reflexiones deductivas, y C. G. Hempel^[4] y Hartley Rogers^[4] han insistido en que un modelo adecuado de la mente tendría que incluir la posibilidad de efectuar reflexiones no deductivas, con las que se obtendría el modo de evitar el resultado de Gödel. Hartley Rogers ha sugerido en concreto que se programe la máquina para tratar varias proposiciones que no hayan sido demostradas ni refutadas, y eventualmente añadirlas a su lista de axiomas. Podrían añadirse, por ejemplo, el último teorema de Fermat o la conjetura de Goldbach. Si posteriormente se viera que su inclusión induce a contradicción, volverían a descartarse y, en ese caso, sus negaciones se añadirían a la lista de teoremas. De este modo podría construirse una máquina capaz de dar por verdaderas ciertas fórmulas que no pudieran demostrarse a partir de axiomas con arreglo a las reglas de inferencia, y, en consecuencia, ya no sería válido el método de demostración de la superioridad de la mente con respecto a la máquina.

Sin embargo, la construcción de una máquina como ésta presenta dificultades. No puede aceptar todas las fórmulas indemostrables añadiéndolas a sus axiomas, lo cual sería incoherente; tampoco serviría que aceptase la primera de cada par de fórmulas indeterminables y, tras añadirlas a sus axiomas, no considerara ya su negación como indeterminable y, en consecuencia, tampoco la aceptara nunca, porque podría suceder en el miembro incorrecto del par la aceptación de la negación de la fórmula gödeliana en lugar de la auténtica fórmula gödeliana. Y el sistema formado por una serie normal de axiomas con la negación de la fórmula de Gödel añadida, aunque no fuera incoherente, sería un sistema poco sólido que no admite interpretación natural. Es algo parecido a las geometrías no desarguianas en dos dimensiones, que no son realmente incoherentes, pero sí lo suficiente incorrectas como para que merezcan

consideración seria. Una máquina susceptible de desafueros de este tipo no sería un modelo de la mente humana.

Está claro que se imponen criterios selectivos bastante claros de fórmulas indemostrables. Hartley Rogers sugiere algunos. Pero, una vez tengamos las reglas que generan nuevos axiomas, aunque los axiomas generados sólo se acepten a título provisional y sean susceptibles de quedar descartados si se comprueba que causan incoherencia, podemos encontrarnos con que hacemos de este sistema, o de cualquier otro un Gödel. Nos encontraríamos como cuando tuvimos la regla que generaba la secuencia infinita de fórmulas gödelianas en forma de axiomas. Es decir, que, independientemente del diseño de una máquina, debe proceder al azar o con arreglo a reglas definidas. En tanto que su procedimiento es aleatorio, no podemos superarla, pero su actuación nunca será un remedo convincente de un comportamiento inteligente; en tanto que su procedimiento responda a reglas definidas, puede utilizarse el método de Gödel para producir una fórmula que la máquina, siguiendo estas reglas, sea incapaz de demostrar, mientras que nosotros, fuera del sistema, comprendemos que es cierta^[5].

El teorema de Gödel sólo es aplicable a sistemas coherentes. Lo único que podemos demostrar *formalmente* es que, *si* el sistema es coherente, la fórmula gödeliana es indemostrable-dentro-del-sistema. Para poder afirmar categóricamente que la fórmula gödeliana es indemostrable-dentro-del-sistema, y por lo tanto verdadera, no sólo hemos de habérmolas con un sistema coherente, sino ser capaces de afirmar que lo es. Y, como demostró Gödel en su segundo teorema —corolario del primero—, es imposible demostrar, en un sistema coherente, que este sistema es coherente. Por lo tanto, para hacer fallar a la máquina produciendo una fórmula de la que pueda afirmarse que es cierta y cuya veracidad no puede demostrar la máquina, tendremos que ser capaces de afirmar que la máquina (o, mejor dicho, su correspondiente sistema formal) es coherente, y no hay prueba definitiva de ello. Lo único que podemos hacer es examinar la máquina y ver si parece coherente. Siempre queda la posibilidad de alguna incoherencia no detectada. En el mejor de los casos, podemos decir que la máquina es incoherente a condición de serlo nosotros. Pero, ¿con qué derecho podemos hacer esto? El segundo teorema de Gödel parece demostrar que una persona es incapaz de afirmar su propia coherencia, por lo que Hartley Rogers^[6] arguye que no podemos realmente utilizar el primer teorema de Gödel para refutar la tesis mecanicista, si no podemos afirmar que «hay atributos inconfundibles que permiten a un ser humano trascender esa limitación afirmando su propia coherencia sin dejar de ser coherente».

La reacción autónoma de una persona, si se pone en tela de juicio su coherencia, es afirmarla con énfasis; pero, en base al segundo teorema de Gödel, algunos filósofos interpretan esto como prueba de su auténtica incoherencia. El profesor Putnam^[8] sugiere que los seres humanos son máquinas, pero máquinas incoherentes. Si se dotara a una máquina de circuitos que mimetizaran un sistema incoherente, no

habría ninguna fórmula bien razonada que no pudiera demostrar, por lo que, en este sentido, no podría demostrarse que fuera inferior a un ser humano. Tampoco podríamos reprocharle su incoherencia, ¿no son también incoherentes las personas? Desde luego lo son las mujeres y los políticos, e incluso los varones que no intervienen en política se contradicen a veces, y una sola incoherencia basta para hacer incoherente un sistema.

No puede negarse el hecho de que a veces seamos incoherentes, pero de ello no se sigue que seamos iguales a sistemas incoherentes. Nuestras incoherencias, más que una actitud fija establecida, son errores equivalentes a la avería accidental de una máquina, no a su esquema normal de operación. Prueba de ello es que evitamos las incoherencias cuando las reconocemos como tales. Si realmente fuésemos máquinas incoherentes, estaríamos tan felices de nuestras incoherencias y afirmaríamos tranquilamente los dos términos de una contradicción. Además, estaríamos dispuestos a decir cualquier cosa, y no es así. Se demuestra fácilmente que, en un sistema formal incoherente, todo es demostrable, y el requisito de coherencia postula precisamente que no todo puede demostrarse dentro de él: no «vale todo». Ciertamente esto es una característica de las operaciones mentales de los seres humanos que son selectivas y dilucidan los enunciados aceptables —verdaderos— de los inaceptables —falsos—, y, cuando una persona está dispuesta a decir cualquier cosa y a contradecirse sin escrúpulos ni reservas, se dice de ella que ha «perdido el juicio». Los seres humanos, aunque no son absolutamente coherentes, son falibles más que incoherentes.

Una máquina falible, pero autocorrectora, seguiría sujeta a los resultados de Gödel; sólo una máquina fundamentalmente incoherente estaría exenta de ellos. ¿Podría obtenerse una máquina fundamentalmente incoherente y al mismo tiempo autocorrectora que estuviera exenta de los resultados de Gödel y al mismo tiempo fuera trivial y totalmente distinta a un ser humano? ¿Una máquina que asumiera una incoherencia lo bastante sofisticada como para ser coherente en todas las operaciones normales, pero capaz de demostrar el axioma godeliano?

Las demostraciones indeseables pueden eludirse de diversas maneras. Podríamos recurrir a una regla según la cual, una vez demostrado p y no- p , se examinarían las pruebas rechazando la más larga. O disponer los axiomas y las reglas de deducción en un orden determinado, y, en el caso de enunciar una prueba que indujera a incoherencia, comprobar los axiomas y reglas en que se basa, descartando el último axioma o regla de ese orden. Siguiendo esta clase de procedimiento podríamos obtener un sistema incoherente con una regla límite, de manera que nunca surgiera la incoherencia en forma de fórmula incoherente.

A primera vista, la sugerencia parece tentadora, pero adolece de un defecto básico. Aunque conservásemos la apariencia de coherencia con una regla mediante la cual, cuando aparecieran dos fórmulas incoherentes, descartásemos la de prueba más larga, nuestro sentido lógico desecharía dicha regla. Aun las sugerencias menos arbitrarias son arbitrarias. Además, el sistema no opera con ciertas reglas deductivas

definidas sobre determinadas fórmulas, sino que las reglas son aplicables y los axiomas ciertos, a condición de que... no lo encontremos incorrecto. Ya nos hemos perdido. Puede aceptarse la regla de Modus Ponens, mientras que otra no; en determinadas ocasiones un axioma puede ser cierto y, en otras, falso en apariencia. El sistema habría tenido que ser un sistema lógico formal, y la máquina difícilmente podría optar a la calificación de modelo de la mente, puesto que distaría mucho de semejarse a la mente en sus operaciones: la mente compara los axiomas dudosos y las reglas de inferencia, pero, si inducen a contradicción, los rechaza de plano. Nosotros procedemos a una comprobación provisional de axiomas y reglas —cierto—, pero no los conservamos si vemos que inducen a contradicción, y los sustituimos por otros si creemos que su formalización es incorrecta y, aunque sea necesario algún axioma o regla de inferencia de esa clase, al no haberlo sabido formular correctamente, no conservamos las formulaciones erróneas sin modificarlas, por la simple razón de que, cuando un razonamiento induce a contradicción, no lo seguimos, ya que hacerlo sería totalmente irracional. Lo idóneo sería que, en las ocasiones en que se nos presentan las premisas de un Modus Ponens, por ejemplo, aplicásemos la regla para extraer la conclusión y, en otras ocasiones, rechazásemos la aplicación de la regla y descartásemos la conclusión. Una persona, o una máquina, que hiciera esto sin ser capaz de aducir un motivo razonable, sería tachada de arbitraria e irracional. El concepto de «argumentos» o «motivos» implica en parte su relativa condición de generales y universales, y, si el Modus Ponens es un método válido para razonar cuando extraigo la conclusión deseada, ha de ser también un método válido cuando mi oponente extrae una conclusión. No quiero aceptarlo. No podemos elegir caprichosamente las ocasiones en que es válida una modalidad de razonamiento si queremos ser razonables. Naturalmente, es cierto que, con nuestras argumentaciones informales, no totalmente formalizadas, diferenciamos entre razonamientos que a primera vista son iguales, añadiendo otros motivos para hacer ver que, sin embargo, lo son; y podría sostenerse que una máquina también puede ser capaz de diferenciar a primera vista razonamientos similares, si tuviera sus buenos motivos para ello. Y también puede sostenerse que la máquina podría tener buenos motivos para rechazar los patrones de razonamiento que rechaza, y por el mejor de los motivos, es decir, para evitar la contradicción. Por eso, si es que constituye un motivo, también es un buen motivo. No atribuimos al mérito de una persona el que evite las contradicciones rechazando simplemente los razonamientos que inducen a ellas por el solo motivo de que, si no, caería en ellas. Este tipo de razonamiento es más un alegato particular que una argumentación sólida. No aumenta el mérito de una persona, con suficiente inteligencia como para prever algunos pasos de la argumentación, cuando trata de no reconocer su propia incoherencia cerrándose en sus trece al ver que el razonamiento se agota. Lo que sucede es que, más bien, consideramos también su incoherencia, no porque en este caso afirme y niegue la proposición misma, sino porque hubiese admitido y rechazado el empleo de la regla deductiva misma. Una regla-límite para

cuando se produce un enunciado incoherente no basta para evitar que se califique de incoherente una máquina incoherente.

Sin embargo, queda la posibilidad de que seamos incoherentes y no haya regla-límite, pero la incoherencia es tan *rebuscada* que nunca aparece. Después de todo, la *ingenua* teoría de los conjuntos, profundamente arraigada en las modalidades reflexivas de sentido común, resultó ser incoherente. ¿Podemos estar seguros de que a la aritmética simple no le espere semejante destino? En cierto sentido, no podemos, a pesar de la profunda sensación de certeza de que nuestro sistema de números enteros susceptibles de adición y multiplicación no vaya a resultar incoherente. Si así fuera, tendríamos que tratar de reformular nuestro concepto intuitivo de número, del mismo modo que el concepto intuitivo de conjunto. Si lo hiciéramos, desde luego, habría que refundir nuestro sistema, los actuales axiomas y reglas de inferencia quedarían radicalmente repudiados y no se plantearía el dilema de utilizarlos y no utilizarlos de un modo «incoherente». Una vez refundido el sistema, estaríamos en las mismas: contaríamos con un sistema que creeríamos coherente, pero indemostrable como tal. ¿Pero no habría también otra incoherencia? Cabe la posibilidad, efectivamente. Pero insisto en que no se tolerarían las incoherencias detectadas. Estamos decididos a no ser incoherentes, dispuestos a erradicar nuestra incoherencia en cuanto aparezca. Por lo tanto, aunque no podemos estar nunca totalmente seguros, ni perfectamente exentos, del riesgo de tener que idear de nuevo las matemáticas, la última posición será una de dos: o disponemos de un sistema de aritmética simple que dentro de nuestras posibilidades creamos coherente, o no hay tal posibilidad de sistema. En el primer caso, estamos en la misma posición que ahora y, en el segundo, si descubrimos que no puede haber ningún sistema basado en la aritmética simple exento de contradicciones, tendremos que abandonar, no ya todas las matemáticas y ciencias anexas, sino toda actividad intelectual.

Aún podría sostenerse que, aunque una persona asuma, en este sentido, su propia incoherencia, no puede en rigor afirmarla sin con ello desmentir sus palabras. Podemos ser coherentes; efectivamente, no existe motivo alguno que nos induzca a pensar que no lo somos, pero una modestia ineludible nos impide afirmarlo. Sin embargo, no es esto precisamente lo que postula el segundo teorema de Gödel. Gödel ha demostrado que, en un sistema coherente, no puede demostrarse dentro *de ese sistema* una fórmula que enuncie la coherencia del mismo. De lo que se sigue que, si una máquina es coherente, no puede demostrar la veracidad del enunciado de su propia coherencia, y, por consiguiente, que una mente, *si realmente fuese una máquina*, no podría llegar a la conclusión de su propia coherencia. Para una mente que no sea una máquina no procede esta conclusión. Lo único que Gödel ha demostrado es que una mente no puede llegar a una prueba formal de la coherencia de un sistema formal dentro del propio sistema, pero no hay objeción a salir del sistema ni a reproducir argumentos informales de la coherencia, ya sea de un sistema formal o de algo menos formal y menos sistematizado. Estos argumentos informales

no serán capaces de formalizarse del todo; por consiguiente, la intención de los resultados de Gödel radica en que no debemos pedir, ni podemos obtener, una formalización completa. Y, aunque hubiera sido agradable obtenerla, como los argumentos totalmente formalizados son más coercitivos que los informales —a pesar de lo cual, nos es imposible integrar en esa modalidad todos los argumentos—, no debemos reprochar a los argumentos informales que sean informales ni considerarlos absolutamente inútiles. Por lo tanto, me parece que, para una mente, es adecuado y razonable afirmar su propia incoherencia: adecuado, porque, aunque las máquinas, como cabía esperar, son incapaces de reflejar totalmente por sí mismas su actuación y capacidad, la capacidad de autoconciencia en este aspecto es lo que se espera de las mentes, y es lo razonable, por los motivos expuestos. No sólo podemos simplemente decir que *sabemos* que somos incoherentes, errores nuestros aparte, sino que, en cualquier caso, debemos *asumir* que lo somos, si pensamos que es posible; además, somos selectivos, no decimos cualquier cosa como lo haría una máquina incoherente; y, finalmente, en cierto sentido, podemos *decidir* ser coherentes, en el sentido de que podemos optar por no tolerar incoherencias en nuestra reflexión y expresión, eliminándolas si surgen, desechando y suprimiendo un extremo de la contradicción.

Así, podemos ver que, tal como posiblemente habíamos supuesto, el teorema de Gödel establece la diferencia entre seres con conciencia propia y objetos inanimados. La esencia de la fórmula gödeliana es su naturaleza autorreferencial; dice que «esta fórmula es indemostrable-dentro-del-sistema». Al trasladarla a una máquina, la fórmula se especifica en términos que dependen de una sola máquina concreta. Se le plantea a la máquina una pregunta sobre sus propios procesos; le pedimos que sea autoconsciente y diga lo que puede y no puede hacer. Son preguntas que evidentemente conducen a una paradoja. Al primer y más mínimo intento de filosofar, se ve uno enredado en preguntas tipo: cuando se sabe algo, se sabe que uno lo sabe, y qué se piensa de ello cuando se piensa sobre uno mismo, y qué efecto causa esta reflexión. Tras sentir la confusión y el peso de este problema durante mucho tiempo, aprende uno a no insistir sobre las preguntas, ya que se entiende, implícitamente, que el concepto de ser consciente es distinto del de objeto inconsciente. Diciendo que un ser consciente sabe algo, decimos no sólo que lo sabe, sino que sabe que lo sabe, y que sabe que sabe qué lo sabe, y así indefinidamente, en tanto nos planteamos la pregunta; reconocemos que esto encierra una infinitud, pero no hay una regresión infinita en el mal sentido, porque las que se agotan son las preguntas, por insustanciales, y no las respuestas. Se intuye que las preguntas son insustanciales porque el concepto encierra la idea de ser capaz de responder indefinidamente a esas preguntas. Aunque los seres conscientes poseen la capacidad de continuar, no se trata de demostrarlo simplemente como una sucesión de tareas que son capaces de realizar, ni tampoco vemos la mente como una secuencia infinita de yos, super-yos y super-super-yos. Más bien se trata de insistir en que un ser

consciente es una unidad y, aunque hablamos de partes de la mente, es tan sólo en sentido metafórico, sin que se entienda al pie de la letra.

Las paradojas de la conciencia surgen porque un ser consciente puede serlo tanto de sí mismo como de otras cosas, sin que por ello se le pueda analizar como divisible en partes. Esto significa que un ser consciente puede hacer frente a las preguntas gödelianas de un modo imposible para una máquina, porque un ser consciente es capaz de considerarse a sí mismo y a su actuación sin ser ajeno a la actuación. Digamos que una máquina puede construirse para que «considere» su propia actuación, pero no puede «tenerla en cuenta» sin convertirse en una máquina distinta, es decir, la máquina original con una «parte nueva» añadida. Pero es inherente a nuestro concepto de mente consciente que se autorrefleje y que critique su propia actuación sin necesidad de parte suplementaria, porque ya es completa y carece de talón de Aquiles.

De este modo la tesis empieza a convertirse en una cuestión de análisis conceptual más que en hallazgo matemático. Lo cual se confirma si consideramos otro razonamiento expuesto por Turing. Hasta la fecha hemos construido artefactos simples y predecibles. Cuando aumentemos la complejidad de nuestras máquinas, quizá nos esperen sorpresas. Turing establece un paralelismo con una pila de fisión: por debajo de un determinado tamaño «crítico» la actividad es insignificante, pero, por encima de ese tamaño crítico, empiezan a saltar chispas. Quizá suceda lo mismo con los cerebros y las máquinas. La mayoría de los cerebros y todas las máquinas actuales son «subcríticos» —reaccionan a los estímulos que les llegan de forma torpe y aburrida, no tienen ideas propias y sólo producen respuestas rutinarias—, pero algunos cerebros actuales, y probablemente algunas máquinas futuras, son supercríticos y chisporrotean por su cuenta. Turing sugiere que es tan sólo cuestión de complejidad y que, por encima de determinado nivel de complejidad, surge una diferencia cualitativa, de modo que las máquinas «supercríticas» serán muy distintas de las sencillas máquinas construidas hasta la fecha.

Puede ser. La complejidad aporta muchas veces diferencias cualitativas. Aunque no parezca plausible, puede que, por encima de determinado nivel de complejidad, una máquina deje de ser predecible, incluso en principio, y empiece a hacer cosas por su cuenta, o, por emplear una expresión reveladora, empiece a tener mente propia al dejar de ser predecible y totalmente dócil, y ser capaz de realizar cosas que consideraríamos inteligentes y no simplemente errores u operaciones al azar, cosas que no hubiéramos programado. Pero entonces dejaría de ser máquina con arreglo al significado de la actuación. Lo que está en juego en el debate mecanicista no es cómo son las mentes, o cómo pueden ser una vez creadas, sino cómo operan. La tesis mecanicista requiere necesariamente que el modelo mecánico de la mente opere con arreglo a «principios mecánicos», es decir que pueda entenderse la operación del conjunto en términos del funcionamiento de sus partes, y que el funcionamiento de cada una de las partes esté determinado por su estado inicial y por la estructura de la

máquina, o que sea una elección aleatoria que opte ante un número determinado de operaciones definidas. Si el mecanicista produce una máquina tan complicada en la que esto no sea aplicable, ya no será una máquina que se ajuste a los términos de la polémica, independientemente de cómo haya sido construida. Cabría más bien decir que ha creado una mente, en el mismo sentido que actualmente se procrean seres. En este caso, habría dos modos de traer nuevas mentes al mundo; la forma tradicional por gestación materna y mediante el nuevo método de construir sistemas, muy, muy complicados, de válvulas y relés, pongamos por caso. Al referirnos al segundo método, debemos insistir en que, aunque lo creado se pareciese a una máquina, no lo sería realmente porque no resultaría el conjunto de sus partes. No podríamos saber lo que iba a hacer por el simple conocimiento de la estructura constructiva y el estado inicial de sus partes; ni siquiera podrían determinarse los límites de lo que podría hacer, porque, incluso al presentarle una pregunta gödeliana, daría la respuesta correcta. De hecho, diremos sucintamente que cualquier sistema que no fuera vencido por la pregunta de Gödel, no sería *eo ipso* una máquina de Turing, es decir, no sería una máquina con arreglo al significado de la actuación.

Si la prueba de la falsedad del mecanismo es válida, resulta de inapreciable repercusión para toda la filosofía. Desde la época de Newton los filósofos han vivido obsesionados por el fantasma del determinismo mecanicista. En estricto sentido científico, habría que considerar a los seres humanos como autómatas determinados y no como agentes morales autónomos; en estricto sentido moral, habría que negar a la ciencia sus derechos, estableciendo un límite arbitrario a sus progresos en el conocimiento de la neurofisiología humana y refugiarnos en el misticismo oscurantista. Ni siquiera Kant pudo resolver la tensión entre ambas perspectivas. Pero actualmente, a pesar de la pervivencia de incontables argumentos en contra de la libertad humana, la tesis mecanicista, quizá la más respetable, ha perdido fuerza. A su sombra no puede ya el filósofo naturalista negar la libertad en nombre de la ciencia, el moralista no puede ya aferrarse a la necesidad de eliminar el conocimiento para dejar sitio a la fe. Incluso podemos comenzar a vislumbrar el modo de hacer sitio a la moral sin necesidad de eliminar o mutilar el enfoque científico. No hemos puesto límites con nuestro argumento a la investigación científica; sigue abierta la puerta del estudio del funcionamiento del cerebro. Sigue siendo posible producir modelos mecánicos de la mente. Sólo que ahora vemos que ningún modelo mecánico nos satisfará plenamente, como tampoco ninguna explicación en términos estrictamente mecanicistas. Podemos producir modelos y explicaciones que sean esclarecedores, pero, por muy lejos que nos lleven, siempre quedará algo por decir. No hay límites arbitrarios a la curiosidad científica, pero ninguna curiosidad científica puede jamás agotar la infinita variedad de la mente humana.

El juego de imitación^[*]

Keith Gunderson

1

Molesto por lo que interpretó como carácter ambiguo, si no absurdo, de la pregunta: «¿Pueden pensar las máquinas?», el fallecido A. M. Turing, en su artículo «Maquinaria computadora e inteligencia» (primer capítulo de este volumen), trató de sustituir del siguiente modo los términos de la pregunta:

«El problema en su nuevo planteamiento puede exponerse en términos de un juego que denominaremos «juego de imitación». Intervienen en él tres personas: un hombre (A), una mujer (B) y un preguntador (C), indistintamente de un sexo u otro. El preguntador se sitúa en una habitación aparte y, para él, el juego consiste en determinar quién de los otros dos es el hombre y quién la mujer. Los conoce por la referencia X e Y, y al final del juego determina si “X es A e Y es B” o si “X es B e Y es A”. El preguntador puede plantear a A y a B preguntas como éstas: “Por favor, X, ¿me dice la longitud de su pelo?”

»Supongamos que X es realmente A, entonces A tiene que contestar. El objetivo de A en el juego es lograr que C efectúe una identificación errónea, por lo que su respuesta podría ser: “Mi pelo es corto, escalonado, y los mechones más largos son de unos veinte centímetros”.

»Para que el preguntador no pueda guiarse por el timbre de voz, las respuestas deben ser por escrito o, mejor aún, mecanografiadas. Lo ideal es disponer de un impresor telegráfico que comunique las dos habitaciones. Otro procedimiento consiste en que un intermediario repita pregunta y respuesta. El objeto del juego para el tercer jugador (B) es ayudar al preguntador. La mejor estrategia para la jugadora es probablemente responder la verdad, añadiendo quizás a sus respuestas cosas como ésta: “¡Soy la mujer, no le haga caso!”, pero de nada sirve, ya que el hombre puede hacer observaciones similares.

»Ahora planteemos la pregunta: “¿Qué sucede cuando una máquina sustituye a A en el juego?” ¿Se pronunciará el preguntador en este caso tan erróneamente como lo hace cuando en el juego participan un hombre y una mujer? Estas preguntas sustituyen a la original: “¿Pueden pensar las máquinas?”»

Y las respuestas de Turing a estas preguntas quedan prácticamente resumidas en el siguiente párrafo: «Personalmente creo que, dentro de unos cincuenta años, podrá

perfectamente programarse computadoras con una capacidad de almacenamiento de 10^9 aproximadamente, para hacerlas jugar el juego de imitación tan bien que un preguntador corriente no tenga más del 70 por ciento de posibilidades de efectuar una identificación correcta a los cinco minutos de plantear preguntas». Y, aunque prosigue reiterando que sospecha que la pregunta original, «¿Pueden pensar las máquinas?», carece de sentido y que debería eliminarse sustituyéndola por una formulación más precisa de los problemas que encierra (una formulación como, por ejemplo, una serie de preguntas sobre el juego de imitación y las capacidades de la máquina), lo que en último extremo se desprende es que, después de todo, Turing contesta a la «absurda» pregunta y que su respuesta es afirmativa, conforme se deduce de sus conclusiones relativas a las capacidades de las máquinas que pueden reemplazar con éxito a las personas en la prueba del juego de imitación.

Hay que señalar que la certidumbre de Turing con respecto a las posibles facultades de las máquinas no se limita a actividades como las de efectuar el juego de imitación tan acertadamente como los seres humanos. No niega, por ejemplo, que sea posible inventar una máquina capaz de apreciar el sabor de las fresas y los helados, aunque considera que sería una «imbecilidad» intentarlo, y en general se limita a establecer una relación positiva de las facultades de las máquinas susceptibles de ejemplificación en términos de su intervención en el juego de imitación.

Por lo tanto, nos centraremos fundamentalmente en contestar si una máquina, capaz de jugar el juego de imitación tan bien como Turing pensaba, es o no es una máquina de la que puede decirse con fundamento que es capaz de pensar, y qué implicaciones encierra semejante afirmación.

Algunos filósofos^[1] están en desacuerdo con el tratamiento que hace Turing de la pregunta «¿Pueden pensar las máquinas?», aunque el juego de imitación, que constituye sin lugar a dudas el núcleo de la conclusión afirmativa, no ha sido objeto más que de simples referencias o alusiones de pasada. Trataré de plantear de una forma más precisa algunas objeciones, objeciones que, personalmente, creo que Turing no supo prever. Por lo tanto, mis observaciones serán fundamentalmente críticas, lo que no quiere decir que no crea que no existen líneas plausibles de defensa para los partidarios de Turing. Por ello cerraré mi argumentación con un escueto esbozo indicativo de cuáles pueden ser algunas y de ciertos interrogantes que, en mi opinión, plantea Turing a los filósofos de la mente, aunque evitaré desarrollarlos.

2

Consideremos la siguiente pregunta: «¿Pueden imitar las piedras?» podría decirse que es una pregunta «demasiado absurda para merecer discusión». Sin embargo, parece factible volver a formular el problema en términos relativamente concretos como los que siguen:

«El problema, en su nuevo planteamiento, puede exponerse en términos de un juego que denominaremos “juego del pisotón”. En él intervienen tres personas: un hombre (A), una mujer (B) y un preguntador (C), indistintamente de un sexo u otro. El preguntador se sitúa en una habitación aparte. La puerta está cerrada, pero, en la pared, a la altura del suelo, hay una pequeña abertura a través de la cual puede introducir casi todo el pie. Cuando lo hace, uno de los jugadores le pisa. Para el preguntador el objeto del juego consiste en determinar por el modo en que le pisan quién de los otros dos es el hombre y quién la mujer. Los conoce por la referencia X e Y, y al final del juego determina si “X es A e Y es B” o si “X es B e Y es A”. Ahora bien, el preguntador —en realidad la persona que recibe el pisotón—, antes de introducir el pie en la abertura, puede señalar si tiene que pisar X o Y. O mejor, puede hacerse una división en la abertura con un lado para X y otro para Y (uno para A y otro para B).

»Supongamos que C introduce el pie por el lado de la abertura correspondiente a A (que en la parte de la pared en que está C puede referenciarse con X e Y). El objeto del juego para A es lograr que C efectúe una identificación errónea. Por consiguiente, el pisotón que dé puede ser rápido y punzante como el de una mujer con zapatos de tacón alto.

»El objeto del juego para el tercer jugador (B) es ayudar a la persona que recibe el pisotón. La mejor estrategia para la jugadora es probablemente intentar pisar del modo más femenino posible. Puede añadir al pisotón matices como el de un ligero movimiento giratorio con el afilado tacón, pero de nada sirve, ya que el hombre puede dar un pisotón igual porque dispondrá de varios modelos de zapato para modificarlo.

»Planteemos ahora la pregunta: “¿Qué sucedería si una caja con piedras (una caja llena de piedras de diverso tamaño, forma y peso), dotada de un ojo eléctrico que actúa a través de la abertura, deja caer una piedra encima del pie de C cuando éste lo introduce por la parte de A, sustituyendo a A en el juego?” (Para hacer la situación más convincente, dotaremos a la caja de piedras de un mecanismo que retire la piedra en cuanto caiga para evitar ruidos delatores como sería el caso si ésta rodara por el suelo, etc.) ¿Se pronunciará el preguntador —la persona que recibe el pisotón— en este caso tan erróneamente como lo hace cuando participan en el juego un hombre y una mujer? Estas preguntas sustituyen a la original “¿Pueden imitar las piedras?”»

Con lo anterior parece demostrarse lo siguiente: lo que se deduce del juego del pisotón no es en absoluto

que las piedras sean capaces de imitar (supongo que nadie optará por esa línea de argumentación), sino tan sólo que se las puede disponer de tal modo que sustituyan al ser humano en el juego del pisotón sin que cambien las características esenciales del juego. Y lo afirmamos, a pesar del hecho de que, si un ser humano fuera capaz de actuar en el juego del pisotón como acabamos de exponer, sería correcto afirmar que esa persona estaría imitando, etc. Por supuesto, una computadora digital es un

mecanismo más excelso que una caja de piedras, pero Turing no ha expuesto argumentación convincente alguna de que su actuación en el juego de imitación, a diferencia de los resultados que logra, sea en modo alguno equiparable a la actuación de un ser humano, del mismo modo que la actuación de la caja de piedras en el juego del pisotón no es equiparable a la actuación de un ser humano. Podríamos llevar demasiado lejos la grotesca paradoja, pero creo que revela perfectamente el motivo por el que no existe contradicción en afirmar: «Sí, una máquina puede jugar al juego de imitación, pero no puede pensar». Por el mismo motivo no existe contradicción si decimos: «Claro que puede hacerse una caja de piedras de tal y tal características, pero las piedras no pueden imitar», pues pensar (o imitar) no pueden describirse inequívocamente con la simple pormenorización de los resultados estrictos citados, ya que, si así no fuera, sería correcto decir que un trozo de tiza piensa o compone porque un vendaval la hizo escribir un rondó sobre una pizarra, o que un fonógrafo canta o una célula fotoeléctrica ve acercarse a la gente.

Se puede impedir la entrada a las personas en un edificio con una célula fotoeléctrica o por medio de un portero. El resultado es el mismo. Pero, aunque un portero puede ser grosero o educado, la célula fotoeléctrica no está a favor ni en contra de la urbanidad. Turing sólo enarbola los resultados estrictos, pero yo creo que lo que acabamos de exponer indica, cuando menos, ciertas dificultades relativas a toda aseveración de pensamiento o decisión sobre si determinado objeto es capaz de un tipo de pensamiento fundamentalmente basado en resultados estrictos. Naturalmente que podría, además, preguntarse si los resultados son realmente iguales en los dos casos, pero no quiero seguir esta línea de argumentación en el debate. Mi interés se centra fundamentalmente en señalar que el aserto de Turing, extraído fundamentalmente de los resultados estrictos, falla por ese motivo. No es una refutación eficaz replicar: «¡Pero parte de los resultados estrictos en cuestión la constituye el engaño de personas inteligentes!», ya que ¿en qué refuerza esto la argumentación general? ¡Claro que la gente puede resultar engañada por cajas de piedras! Es de dominio público que los fonógrafos de alta fidelidad han alcanzando tal nivel de perfección que los críticos con los ojos vendados son incapaces de distinguir su «interpretación» de la del Cuarteto de Cuerdas de Budapest, por ejemplo. Pero nunca se dirá que el fonógrafo interpretó el sábado con inusitada brillantez, ni se le pedirá un bis.

3

Puede que unos resultados estrictos equiparables, obtenidos por máquinas y seres humanos, basten para establecer una analogía entre los dos, pero, aun así, distan mucho de darse las condiciones necesarias para establecer que una clase de sujeto (las máquinas) hace lo mismo que hace otra clase de sujeto (los seres humanos u otros animales). Y parte de las cosas que hacen es el modo cómo las hacen. Preguntar si

una máquina puede pensar forma parte de si las máquinas pueden hacer las cosas de cierto modo.

Lo anterior es relevante en lo que podríamos denominar el problema de la diferenciación y evaluación de los resultados estrictos que obtiene una máquina, tal como lo esboza Scriven en la discusión de lo que él denomina «el problema actuatorio» y «el problema de la personalidad». Escribe este autor en «The Compleat Robot: A Prolegomena to Andriology» (cf. bibliografía):

«El problema de cómo actúa, en este caso, es el de si una computadora puede obtener resultados que, transcritos, nos faciliten lo que pudiéramos calificar de solución o prueba original, *si procediera de una persona*. El problema de la personalidad estriba en si a este resultado se le puede denominar solución o prueba, a pesar de que *no* procede de una persona».

Y prosigue:

«La trampa lógica es ésta: ningún resultado de actuación basta para persuadirnos a emplear el vocabulario correspondiente al de comportamiento humano, pero si no admitimos emplear este léxico en cada caso particular, en base a este razonamiento, quizá nos hayamos comprometido, erróneamente, a eludirlo aun en el caso de que se obtengan simultáneamente *todos* los comportamientos».

Sin embargo, no me preocupa lo que se haya de calificar de solución o prueba original. Scriven, en los citados párrafos se refiere a las afirmaciones: «Las máquinas sólo hacen lo que nosotros les ordenemos. Son incapaces de auténtico pensamiento original», y comenta que «se juntan dos cuestiones de diferencia muy notable». Precisamente con lo expuesto anteriormente se intenta dilucidar estas cuestiones, pero, al parecer, hay incluso tres y no dos cuestiones unidas en las citadas afirmaciones. La tercera, no tratada por Scriven en su diferenciación del problema de la actuación y el de la personalidad, es sencillamente el problema mencionado anteriormente: el de discernir cuándo un sujeto (una máquina) ha *hecho lo mismo* que otro sujeto (un ser humano). Y por «hacer lo mismo» no se entiende «lograr igual resultado final» sin más. (Lo cual no equivale a sugerir que no pueda nunca emplearse la expresión en relación con el acto de pensar). Esto es importante en relación al análisis de Scriven, ya que podría darse el caso de que una máquina obtuviera simultáneamente todos los resultados, como sugiere Scriven, y que, en base a diversos razonamientos, decidiéramos calificarlos de pruebas y soluciones originales, superando con ello el problema de la personalidad, aunque aún siguiéramos reticentes en cuanto a admitir que las máquinas fueran capaces de «pensamiento auténtico original». Los fundamentos de esta decisión podrían equipararse perfectamente a los fundamentos en que nos apoyaríamos para decir que las piedras pueden imitar (a pesar de que las cajas de piedras hubieran alcanzado un

alto nivel tecnológico). Desde luego, podríamos basarnos en razonamientos no tan sólidos, pero me limito a imaginar el caso en que lo fueran, que es, además, el caso en que se obtienen todos los resultados, de tal forma que pueden considerarse soluciones o pruebas originales. En esta tesitura veríamos que las respuestas a las preguntas sobre originalidad y actuación, y la trampa lógica mencionada por Scriven, deben ir totalmente diferenciadas en cualquier respuesta que pudiera darse a la pregunta de si las máquinas implican o no pensamiento, por lo que no serían respuestas adecuadas a la pregunta de si son o no capaces de «auténtico pensamiento original». En otras palabras, son distintas las preguntas sobre la originalidad y sobre el pensamiento, pero esta diferencia no se recoge en el razonamiento de Scriven.

4

Pero volvamos al juego de imitación. Es de suponer que, si los seres humanos participaran en este juego, casi con toda seguridad imaginaríamos que deliberarían, decidirían y reflexionarían —es decir, que «se pensarían las cosas» conforme se intercambiaran los mensajes. Y, si alguien nos preguntara, por ejemplo, sobre la capacidad intelectual o la habilidad mental de Johnson, podríamos citar ese juego que él solía jugar y cómo derrotaba a Peterson y a Hanson que jugaban con él. Pero no lo consideraríamos más que un simple ejemplo posible de la capacidad mental de Peterson, y en general no nos veríamos en aprieto alguno al citar incontables ejemplos de la capacidad de Peterson para deliberar, imaginar, extrañarse, reflexionar, o lo que, resumiendo, podemos llamar pensamiento. Podríamos, por ejemplo, explicar cómo compone sonetos o cómo discute con Hanson. Ahora bien, no es que desee afirmar que esté fuera del alcance de una máquina hacer todo esto; en realidad no intento, en este artículo, dar una respuesta a la pregunta «¿Pueden pensar las máquinas?»», pero lo que sí quiero señalar es que lo que digamos sobre Peterson en otra serie interminable de situaciones tenderá a influir en lo que digamos de él a propósito del juego de imitación. Una piedra rueda por la ladera de una montaña y, de hecho, no hay comportamiento o actividad por su parte. Pero, si un hombre rueda por un monte, podemos con toda justicia preguntarnos si le empujaron o lo hizo intencionadamente, si se divierte jugando, o simulando ser una planta rodadora, o qué. No podemos limitarnos a pensar que una persona rueda por un monte, a menos que esté muerta. *A fortiori* no podemos entender que participe en el juego de imitación, al margen de sus disposiciones, hábitos, etc., por los que se destaca en contextos distintos al del juego de imitación. Por lo tanto, no cabe esperar que hallemos una respuesta definitiva al problema de cómo categorizar una máquina que puede jugar (bien) al juego de imitación. El pensamiento, independientemente de que cualquier categorización o calificación positivas a él referidas sean correctas, no es algo que se explica o define con un ejemplo cualquiera. Sin embargo, la parte del

planteamiento de Turing a que me refiero se fundamenta ostensiblemente en un solo ejemplo.

5

Lo que sigue puede ayudar a esclarecer lo expuesto. Imaginemos el siguiente diálogo:

«*Vendedor de aspiradoras*: Aquí tiene un ejemplo de lo que la Swish-600-todo-uso puede hacer. (Acto seguido aplica la boquilla a la alfombra y aspira una mota de polvo.)

»*Ama de casa*: ¿Y qué más puede hacer?

»*Vendedor de aspiradoras*: ¿Qué quiere usted decir con “¿Qué más puede hacer?” Ha aspirado esa mota de polvo, ¿no lo ha visto?

»*Ama de casa*: Sí, ya he visto que aspira una mota de polvo, pero pensaba que era todo-uso. ¿No aspira trozos más grandes y pesados de paja, papel o barro? ¿Y los pelos de gato del sofá?

»*Vendedor de aspiradoras*: Aspira motas de polvo. Las aspiradoras son para eso.

»*Ama de casa*: ¡Ah! eso es lo que hace. Pensé que sólo era un ejemplo de lo que hacía.

»*Vendedor de aspiradoras*: Es un ejemplo de lo que hace. Lo que hace es aspirar motas de polvo».

Preguntemos: ¿Quién tiene razón en cuanto a los ejemplos? Respondemos: no está del todo claro que nadie mienta o utilice indebidamente la palabra «ejemplo». Y no hay una regla lingüística clara ni una reglamentación que nos indiquen que, si S sólo puede hacer x, el hecho de que S haga x no pueda constituir un ejemplo de lo que S puede hacer, ya que el que sea un ejemplo presupone, o implica, o qué sé yo, que hay toda una serie de otros ejemplos (aspirar barro, pelos de gatos, etc.). Sin embargo, a pesar de ello, el ama de casa tiene razón. Uno tiene perfecto derecho a esperar más de una Swish-6(X)-todo-uso de lo que ha visto en la demostración. Está claro que aquí el principal problema radica en «todo-uso» más que en «ejemplo», aunque también pueda existir cierto confucionismo con respecto a «Aquí tiene usted un ejemplo...», y, desde luego, sería inducir a error decir: «Esto es *un simple ejemplo*», seguido de «...de lo que la Swish-6(X)-todo-uso puede hacer». La relevancia filosófica de todo ello dentro de nuestro planteamiento puede resumirse de la siguiente forma, más bien casera: «pensar» es un término que comparte ciertas características de «todo-uso», tal como figura en la expresión «Swish-600-todo-uso». No se emplea para designar o referirse a una capacidad, habilidad, disposición, talento, hábito o característica de un sujeto determinado, igual que tampoco «todo-uso» se emplea en el ejemplo para destacar ninguna operación concreta de la

aspiradora. Pensar, independientemente de la atribución positiva que se le pueda dar, no es, pongamos por caso, como nadar o jugar al tenis. La cuestión de si Peterson sabe nadar o jugar al tenis puede resolverse con unos cuantos ejemplos-prueba de Peterson nadando o jugando al tenis. (Y se observará que es difícilmente imaginable que se plantee la cuestión de si Peterson puede o no pensar. Ya que en general no es nada interesante plantearla a propósito de nuestros prójimos, si bien sería interesante que nuestros prójimos se la planteasen con respecto a los diversos antropoides considerados en las múltiples fases de la evolución). Pero, si suponemos que se plantea la cuestión refiriéndose a Peterson, la única clase de respuesta adecuada sería algo así: «¡Por Dios!, ¿qué le hace pensar que no puede?» (como si temiéramos la noticia de alguna terrible lesión cerebral en Peterson). Y no nos sorprendería que hubiera quizá perdido una facultad concreta. No sería como el caso de que el campeón de Wimbledon hubiera perdido sus cualidades de jugador por amputación del brazo.

No es más extraordinaria para un ser humano la facultad de pensar que para un ser humano estar compuesto de células. Igualmente, «Puede pensar» es tan poco una respuesta a preguntas relativas a la facultad mental o inteligencia de Peterson como lo es «Está compuesto de células» a la habitual pregunta acerca del aspecto exterior de Peterson. Y decir que Peterson puede pensar no es decir que otros ejemplos prácticos del acto de pensar van inscritos en sus huellas dactilares, como tampoco decir que la Swish 600 es todo-uso es suponer una o dos operaciones concretas de las que realiza el aparato. Porque el acto de pensar no puede identificarse con lo que se demuestra en un ejemplo o una clase de ejemplo concreto; por lo tanto, el enfoque de Turing de responder a la pregunta «¿Puede pensar una máquina?» con el juego de imitación deja mucho que desear. Efectivamente, nos sugiere un diálogo muy parecido al anterior:

«*Turing*: ¿Sabe usted una cosa? Las máquinas pueden pensar.

»*Filósofo*: ¡Santo cielo! ¿De verdad? ¿Cómo es posible?

»*Turing*: Pues porque pueden jugar al juego llamado de imitación (sigue una explicación del juego).

»*Filósofo*: Interesante. ¿Qué más pueden hacer? Podrán hacer muchas cosas si realmente piensan.

»*Turing*: ¿Qué quiere usted decir con “¿Qué más pueden hacer?” Juegan al juego de imitación. Eso es pensar, ¿no?

»Etc.».

Pero Turing, como el vendedor de aspiradoras, tiene dificultades para vender. De todos modos, paso a señalar brevemente por qué hay que modificar algunas de nuestras críticas a su enfoque.

Haremos una última observación crítica antes de señalar ciertos defectos de la exposición. Como hemos señalado, el argumento de Turing se beneficia al poner de relieve el hecho de que una máquina sustituye a un ser humano en una situación concreta y actúa tan bien como actuaría el ser humano en esa situación. Sin embargo, nadie negará que las máquinas son capaces de hacer determinadas cosas igual de bien, o mejor, que los seres humanos, por lo que nadie dirá que ninguno de estos casos constituye de por sí un argumento a favor de la afirmación de que las máquinas pueden pensar. Ya que, en muchos de estos casos, podría señalarse, en lugar de insistir en que una máquina puede hacer lo mismo que un ser humano, el hecho de que apenas se requiere ser una persona para hacer estas cosas. Por ejemplo: «En mi trabajo ni siquiera tengo que pensar; me limito a precintar los tarros conforme pasan en la cinta transportadora», o: «Sirvo refrescos, uno tras otro, como una máquina». Difícilmente se interpretaría esta última afirmación como: «¡Dios mío, qué listas son las máquinas que despachan refrescos!», sino más bien como: «¡Qué idiotez es mi trabajo!, no requiere esfuerzo mental alguno». Además, como me ha sugerido el profesor Ryle, un cajero veterano de banco es capaz de sumar, restar, multiplicar y dividir sin esfuerzo mental mientras piensa en otras cosas, y ¿no recitamos nosotros el abecedario o una canción conocida sin necesidad de pensar? No pretendo que esto sea una crítica específica a Turing; sólo trato de recordar que ser capaz de hacer lo que hacen los seres humanos difícilmente supone que se posean facultades intelectuales o mentales, reales o imitativas, ya que muchas de las cosas que realizan los seres humanos implican poca o nula reflexión. Los parados constituyen una parcela demográfica distinta de la de los que no tienen ingenio.

De todas formas, las siguientes consideraciones atenúan algunas de las críticas expuestas. Los partidarios de Turing pondrían de relieve que una máquina capaz de jugar el juego de imitación es también capaz de hacer mucho más: puede computar, quizá ser programada para jugar al ajedrez, etc. y, en consecuencia, muestra facultades que superan las que puedan exponerse en «el ejemplo» sobre el que hemos insistido en nuestras críticas. No entraré en los detalles que considero réplica adecuada a este alegato, pero en general creo que sería posible formular una refutación, siguiendo el mismo razonamiento, que demostrase que jugar al ajedrez, calcular y realizar otras operaciones (fundamentalmente de índole computacional) constituyen una serie de ejemplos, cuando menos limitada, que seguiría sin satisfacer el concepto del acto de pensar. El caso paralelo al del Ama de casa y el Vendedor de aspiradoras sería aquél en el que el Ama de casa siguiera sin aceptar la aspiradora en

su calidad de «todo-uso» a pesar de que le hubieran demostrado su capacidad de recoger restos más consistentes que simples motas de polvo. No obstante, aunque nuestra refutación fuese satisfactoria, seguiría sin responder a la cuestión más general: ¿Qué clase de ejemplos satisfaría los requisitos implícitos de que nos servimos para calificar a los sujetos como «capaces de pensar»?

Corolario: si queremos mantener el interés de la pregunta «¿Pueden pensar las máquinas?», no podemos pretender dar una respuesta positiva sobre la simple base de que la máquina no remeda la actividad humana en todos los aspectos. La pregunta «¿Pueden pensar las máquinas, si pueden hacer lo que hace un ser humano?» no es una pregunta interesante, aunque no deje de ser interesante plantearse si no sería una contradicción lógica suponer que algo así fuera de hecho una máquina. Pero, mientras pensamos en sujetos que son sin lugar a dudas máquinas, hay que estar dispuestos a renunciar del todo a que sus actividades mimeticen completamente los actos humanos para afirmar que piensan, si es que pueden. Pero del todo, ¿hasta qué extremo? Vuelve a plantearse la citada serie de preguntas con respecto a la variedad y a la amplitud de ejemplos necesarios.

Además, puede rebatirse que, dado el creciente papel que desempeñan las máquinas en nuestra sociedad, hasta la palabra «pensar» adquiere nuevos significados y que no es insensato suponer que cambie de significado hasta el punto de que, dentro de cincuenta años, a una máquina que juegue el juego de imitación se la denomine, en la jerga habitual, máquina pensante. Sin embargo, hay una diferencia entre preguntarse si una máquina puede pensar en el contexto de los significados y empleos corrientes de los vocablos «máquina» y «pensar», y el preguntarse si una máquina puede pensar dados los cambios de sentido de esos vocablos. En este trabajo he centrado mi atención en la primera pregunta. No obstante, existe una duda transitoria en la pregunta «¿Pueden pensar las máquinas?», ya que, si se plantea con respecto a un futuro previsible, puede resultar difícil discutir ese futuro sin especificar el cambio de significado y del empleo de la palabra. Pero esto plantea ya otro tipo de perspectiva que la falta de espacio nos impide discutir. Hasta cierto punto los planteamientos de Turing se basan en ciertas convicciones personales respecto de cómo hablaríamos de las máquinas en el futuro, convicciones que nunca expuso detalladamente, aparte de que tampoco trató directamente el espinoso problema de significados relacionados con semejante posibilidad.

Una última puntualización: a veces se adopta la postura de que pensar es la capacidad suprema de la raza humana y que, al negar que las máquinas puedan pensar, lo que se hace es postergarlas a un nivel de actuación inferior al de los seres humanos. Pero podría perfectamente sostenerse que las máquinas no pueden pensar, ya que hacen algo mucho mejor. Podríamos negar hasta la saciedad que las máquinas

puedan pensar en el contexto de un problema matemático, y seguir afirmando que en muchos aspectos su capacidad es de un nivel superior al alcanzado por los seres pensantes, ya que las máquinas pueden dar casi instantánea e infaliblemente respuestas exactas y a veces originales a muchos problemas matemáticos complejos e intrincados, pero no tienen que «pensarse» las respuestas. La perforadora de vapor derrotó definitivamente al excavador de túneles John Henry, pero eso no demuestra que la máquina tuviera músculos; únicamente que no eran imprescindibles los músculos para excavar túneles.

Mentes y máquinas^[*]

Hilary Putnam

Las diversas soluciones y los múltiples dilemas que constituyen el conocido problema mente-cuerpo son exclusivamente de índole lingüística o lógica, y los pocos «hechos» empíricos en este terreno confirman equilibradamente ambas perspectivas. Con este trabajo no pretendo sentar cátedra, pero sí confío en dar un paso que haga más plausibles mis argumentos. Trataré de demostrar en concreto que todas las soluciones se orientan hacia un sistema de computación capaz de responder preguntas sobre su propia estructura, y, por lo tanto, nada tienen que ver con la naturaleza única (si es que es única) de la experiencia subjetiva humana.

Valga, a título de ejemplo, un rompecabezas al que a veces se hace referencia al hablar del «problema mente-cuerpo»: el rompecabezas de la *privacidad*. La pregunta «¿Cómo sé que me duele?» es una pregunta *desviada*^[1] («lógicamente rara»). Esta diferencia puede también detectarse en preguntas impersonales: «¿Cómo sabe realmente alguien que le duele algo?» es desviada; «¿Cómo sabe realmente uno que le duele a otra persona?» no es desviada. Demostraré que se produce una réplica exacta de la diferencia de categoría entre las dos últimas preguntas en el caso de las máquinas: si T es una *máquina de Turing* cualquiera (cf. más adelante), la pregunta «¿Cómo verifica T que está en el estado A?» es a todas luces, como veremos, «lógicamente rara»; pero, si T es capaz de explicar una máquina anexa T' (pongamos por caso que T posee «órganos sensoriales» electrónicos que «barren» a T'), la pregunta «¿Cómo verifica T que T está en el estado A?» no es nada rara.

Otra pregunta relacionada con el «problema mente-cuerpo» es la de si de alguna manera se puede o no identificar los acontecimientos mentales con los acontecimientos físicos. Naturalmente, no pretendo afirmar que esto se dé en el caso de las máquinas de Turing, pero sí digo que es posible construir un análogo lógico para la pregunta que se plantea, y que todos los argumentos que se esgrimen desde las dos vertientes de la cuestión de la «identidad mente-cuerpo» pueden reflejarse en los términos fijados por el análogo.

Para obtener semejante análogo, identifiquemos una teoría científica con un «cálculo parcialmente interpretado» en el sentido que le da Carnap^[2]. Después podemos perfectamente imaginar una máquina de Turing que genere teorías, las verifique (suponiendo que sea posible «mecanizar» hasta cierto punto la lógica inductiva) y «acepte» las teorías que cumplan determinados requisitos (por ejemplo,

el éxito de la predicción). Si, en concreto, la máquina posee «órganos sensoriales» electrónicos que la facultan para «autobarrerse» durante la operación, formulará teorías sobre su propia estructura y las someterá a verificación. Supongamos que la máquina se halla en un estado determinado (pongamos el «estado A») cuando, y sólo cuando, está actuando el *flip-flop* 36. En tal caso, la afirmación: «Estoy en el estado A cuando, y sólo cuando, está actuando el *flip-flop* 36» puede ser uno de los principios teóricos relativos a su propia estructura que la máquina acepta. Naturalmente, en este caso, «Estoy en el estado A» es «lenguaje de observación» para la máquina, mientras que «está actuando el *flip-flop* 36» es una «expresión teórica» que parcialmente se interpreta en términos de «observables» (si los «órganos sensoriales» de la máquina informan mediante símbolos impresos sobre la cinta de entrada de la misma, los «observables», en cuyos términos la máquina da una definición operacional parcial de «actuando el *flip-flop* 36», sería algo así como «símbolo # tal cual va apareciendo en la cinta de entrada»). Ahora bien, todas las consideraciones habituales a favor y en contra de la identificación mente-cuerpo pueden parangonarse con las consideraciones a favor y en contra de la afirmación de que el estado A es, de hecho, *idéntico* a la actuación del *flip-flop* 36.

Los argumentos de «identidad» son *occamistas* tanto en un caso como en otro, y el argumento habitual de dualismo en el caso mente-cuerpo puede equipararse al que se aplica a la máquina, ya que para ésta «el estado A» es directamente observable y, por el contrario, sólo conoce los *flip-flops* mediante inferencias altamente elaboradas. ¿Cómo *podría* jamás ser posible que dos cosas tan distintas sean iguales?

Podemos dar a esta argumentación una forma que sirva para dotarla de una apariencia algo más sólida.

(1) Estoy en el estado A si, y sólo si, está actuando el *flip-flop* 36, es claramente una propuesta «sintética» para la máquina. Por ejemplo, la máquina puede estar en el estado A y sus órganos sensoriales informar que *no* actuaba el *flip-flop* 36. En tal caso, la máquina tendrá que efectuar una elección «metodológica» —o sea para descartar (1) o concluir que había incurrido en un «error de observación» (del mismo modo que un científico se enfrenta necesariamente a elecciones metodológicas al estudiar sus propias correlaciones psicofísicas). Igualmente, los filósofos que discuten a partir de la naturaleza sintética de la proposición:

(2) Me duele si, y sólo si, se estimulan mis fibras C, llegan a la conclusión de que las *propiedades* (o los «estados» o los «acontecimientos») sentir dolor y el estímulo de las fibras C posiblemente no sean lo mismo [de lo contrario (2) sería analítica, o al menos así lo indica la argumentación]; de modo que se podría concluir, a partir del hecho de que (1) es sintética, que las dos propiedades (o «estados» o «acontecimientos») —estar en el estado A y tener en acción el *flip-flop* 36— ¡tampoco son posiblemente lo mismo!

Conviene señalar que el argumento tradicional sobre el dualismo no es en absoluto una conclusión a partir de «los datos crudos de la experiencia directa» (como lo demuestra el hecho de que es igualmente aplicable a máquinas insensibles), sino un razonamiento muy enrevesado que depende de A) la reificación de universales^[3] (por ejemplo, «propiedades», «estados», «acontecimientos») y B) de una notable diferenciación analítico-sintética.

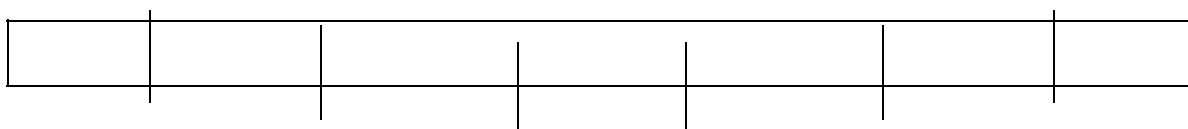
Tal vez se me acuse de proponer una visión «mecanicista» del mundo al extremar semejante analogía. Si esto deja presuponer que sostengo la opinión de que las máquinas piensan^[4], por un lado, y de que los seres humanos son máquinas, por otro, se trata una acusación infundada. Si existe una versión lo bastante sofisticada del mecanicismo como para disipar semejante error, no cabe duda de que los argumentos que se exponen en este trabajo no pretenden sino contribuir a reafirmarla^[5].

1. Máquinas de Turing

Este trabajo requiere la utilización del concepto máquina de Turing^[6] que pasamos a explicar.

Diremos brevemente que una máquina de Turing es un dispositivo con un número finito de configuraciones internas, cada una de las cuales presupone que la máquina está en uno de un número finito de *estados*^[7], y que la máquina explore una cinta que lleva inscritos ciertos símbolos.

La cinta de la máquina está dividida en casillas, o posiciones, como éstas:



y, en cada una de ellas, hay impreso un símbolo (de un determinado alfabeto finito). La máquina posee también un «traceador» que «barre» una por una las posiciones de la cinta. Finalmente, la máquina está dotada de un *mecanismo de impresión* por el que (A) *borra* el símbolo de la casilla barrida y (B) *imprime* en ella otro símbolo (de su alfabeto).

Cualquier máquina de Turing se ajusta totalmente a la descripción de la *tabla de instrucciones*, y su estructura es la siguiente: las filas de la tabla corresponden a letras del alfabeto (incluida la letra «nulo», es decir, espacio en blanco) y las columnas corresponden a estados, A, B, C, etc. En cada posición aparece una «instrucción», por ejemplo, «s₅L A», «s₇C B», «s₃R C». Estas instrucciones se leen del modo siguiente: «s₅L A» significa «escribe el símbolo s₅ en la casilla que estás barriendo (tras borrar

el símbolo que contenga) y procede a barrer la casilla inmediata de la izquierda a la que acabas de barrer; luego cambia al estado A». Las otras instrucciones se interpretan de modo similar («R» significa «barre la casilla inmediatamente a la derecha, y «C» significa «centro», es decir, continúa barriendo la *misma* casilla).

Ejemplo de tabla de máquina:

		A	B	C	D
(s ₁)	1	s ₁ RA	s ₁ LB	s ₃ LD	s ₁ CD
(s ₂)	+	s ₁ LB	s ₂ CD	s ₂ LD	s ₂ CD
	espacio				
(s ₃)	en blanco	s ₃ CD	s ₃ RC	s ₃ LD	s ₃ CD

La máquina descrita según esta tabla está estructurada para funcionar del siguiente modo: empieza a funcionar en el estado A. En la cinta aparece una «suma» (en notación unitaria) a «efectuar», por ejemplo, «11 + 111».

En principio la máquina barre el primer «1». A continuación, procede a «efectuar» la suma (fundamentalmente sustituyendo el signo más por un 1 y, a continuación, retrocediendo y borrando el primer 1). Por lo tanto, si la «entrada» era 1111 + 11111, la máquina «escribe» 111111111 y luego pasa al «estado de reposo» (estado D).

Una «tabla de instrucciones» *describe* la máquina si ésta posee estados internos correspondientes a las columnas de la tabla y si «obedece» a las instrucciones de la tabla en el siguiente sentido: cuando está barriendo una casilla en la que figura el símbolo s₁ y está, pongamos por caso, en el estado B, es decir que sigue la «instrucción» en la fila y columna apropiadas de la tabla (en este caso, la columna B y la fila s₃). Cualquier máquina descrita por una tabla de instrucciones de este tipo es una máquina de Turing.

El concepto de máquina de Turing se halla igualmente sujeto a generalización^[8] en diversas modalidades; por ejemplo, podemos suponer que la máquina posee una segunda cinta (una «cinta de entrada») en la que puede imprimir información suplementaria durante la computación un operador. A continuación, emplearemos esta generalización (y los «órganos sensoriales» electrónicos sustituirán al «operador»).

Hay que señalar que las máquinas de Turing son, en principio, capaces de hacer lo que haga^[9] cualquier computadora (del tipo que sea).

En ciertas ocasiones se ha puesto en tela de juicio (por ejemplo, en *Gödel's Proof* de Nagel y Newman) el que el teorema (de Gödel) indique que la estructura y la potencia de la mente humana son con gran diferencia más complejas y sutiles que cualquier máquina inexistente, por inventar (pág. 10), y que, en consecuencia,

ninguna máquina de Turing puede servir de modelo a la mente humana; pero esto es sencillamente un error.

Supongamos que T sea una máquina de Turing que me «representara», en el sentido de que T pudiera demostrar los postulados matemáticos que yo demostrara. En tal caso, la argumentación (Nagel y Newman no exponen una argumentación, pero supongo que deben pensar en ésta) consiste en decir que, empleando la técnica de Gödel, puedo descubrir una propuesta que T sea incapaz de demostrar, y además yo puedo demostrar esta propuesta. Con esto se refuta el que T me «representa»; por lo tanto, yo no soy una máquina de Turing. El sofisma es una aplicación errónea pura y simple del teorema de Gödel. Dada una máquina arbitraria T, lo único que puedo hacer es hallar una propuesta U con la que yo pueda demostrar:

(3) que, si T es coherente, U es cierta, en la que U es indeterminable por T, si T es realmente coherente. Sin embargo, ¿T puede también demostrar perfectamente (3)! Y la afirmación U, que T *no puede* demostrar (asumiendo la coherencia), tampoco yo puedo demostrarla (a menos que pueda demostrar que T es incoherente, ¡lo cual es improbable, si T es muy complicada!)

2. Privacidad

Supongamos que una máquina de Turing esté construida para lo siguiente: se escribe una cifra, por ejemplo «3000», en la cinta de T y ésta empieza a funcionar en su «estado inicial». A partir de ello, T computa el trimilésimo dígito (o la cifra que fuese) en la expansión decimal de π , escribe el dígito en la cinta y pasa al «estado de reposo» (es decir, se desconecta sola). Está claro que la pregunta «¿Cómo “verifica” (o “computa” o “resuelve”) T el trimilésimo dígito en la expansión decimal de π ?» es aceptable, y la respuesta puede ser muy complicada. De hecho, la respuesta implicaría probablemente tres distintos apartados:

a) Descripción de la secuencia de estados por los que pasa T para dar el resultado, y lectura de la cinta en cada fase de la computación.

b) Descripción de las *reglas* por las que se rige T (según la «tabla de instrucciones» específica de T).

c) Explicación de la *lógica* de todo el procedimiento.

Supongamos ahora que alguien plantea la siguiente objeción: «Para efectuar la computación descrita, T debe pasar por los estados A, B, C, etc. Pero ¿cómo puede T verificar que se halla en los estados A, B, C, etc.?»

Está claro que es una objeción tonta. Pero ¿en qué es tonta? Para empezar, la «descripción lógica» (tabla de instrucciones) de la máquina describe los estados únicamente en términos de sus *relaciones* recíprocas y de lo que aparece en la cinta. La «actuación física» de la máquina es inmaterial en tanto *haya* estados distintos A, B, C, etc., y éstos se sucedan, con arreglo a lo especificado en la tabla de instrucciones. Por lo tanto, se puede contestar una pregunta como «¿Cómo verifica T

a X?» (o «computa», etc.) sólo en el sentido de describir la *secuencia de estados* por los que tiene que pasar T para resolver X (computar X, etc.), las reglas que sigue, etc. ¡Pero T no tiene que pasar por una «secuencia de estados» para estar en un solo estado! Supongamos, efectivamente, que los hubiera; supongamos que T *no* pudiera estar en el estado A sin *verificar* primero que está en el estado A (pasando previamente por una secuencia de estados). Está claro que esto implicaría una regresión infinita. Interrumpimos esta «regresión» sencillamente señalando que la máquina, al comprobar el trimilésimo dígito en π , *pasa por* sus estados, pero no necesita «verificar» en un sentido significativo que pase por ellos.

«Obsérvese la analogía con un sofisma de la epistemología clásica: el de suponer que, para conocer p (siendo p una proposición), hay primero que conocer q_1 , q_2 , etc. (siendo q_1 , q_2 , etc., *otras* proposiciones al caso). Esto nos conduce a una «regresión infinita» o al dudoso formulismo de inventar un «protocolo» especial de proposiciones.

»El sofisma se resuelve de forma análoga a la del caso de la máquina. Supongamos que, en base a experiencias sensoriales E_1 , E_2 , etc., sé que hay una silla en la habitación. De ello no se sigue que verbalice (o incluso *pueda* haber verbalizado) E_1 ; E_2 , etc., ni que recuerde E_1 , E_2 , etc., ni siquiera que «clasifique mentalmente» («ponga atención», etc.) las experiencias sensoriales E_1 , E_2 , etc., cuando las tengo. En resumen, es necesario *tener* experiencias sensoriales, pero no *saber* (ni siquiera *advertir*) qué experiencias sensoriales se tienen para poseer ciertos tipos de conocimiento».

Sin embargo, modifiquemos el caso y supongamos que siempre que una máquina está en un estado particular (pongamos el «estado A»), escribe las palabras «Estoy en el estado A». Luego alguien nos asegura que la máquina en general verifica el estado en que se halla, pero puede decirnos en el caso del estado A (después que la máquina ha escrito «Estoy en el estado A»): «La máquina ha verificado que estaba en el estado A».

Estudiemos este caso con más detalle. Primero, queremos suponer que, cuando se halla en el estado A, la máquina escribe «Estoy en el estado A» sin pasar previamente por ningún otro estado. Es decir, en cada línea de la columna de la tabla señalada con «estado A» aparece la instrucción: *escribe*^[10] «Estoy en el estado A». Segundo, a título comparativo, supongamos un ser humano. Jones, que dice: «Tengo dolor» (o «¡Ay!» o «Me duele») siempre que siente dolor. Para que la comparación sea lo más estricta posible, tendremos que suponer que el condicionamiento lingüístico de Jones es tal que se limita a decir: «Tengo dolor», «sin pensar», es decir, sin pasar por estado mental de introspección alguno al margen del propio dolor. En la terminología de Wittgenstein, Jones se limita a *mostrar* su dolor diciendo: «Tengo dolor»; no lo refleja primero (o presta atención, o lo nota, etc.) para luego describirlo

conscientemente. (Obsérvese que los epistemólogos tradicionales, desde Hume a Russell, pasaron por alto esta simple posibilidad de enunciar la proposición «Tengo dolor» sin previamente realizar «acto de juicio» mental alguno). Consideremos ahora las preguntas paralelas: «¿“Verifica” la máquina que se halla en el estado A?», «¿“Sabe” Jones que tiene dolor?» y sus consecuencias.

Los filósofos interesados por las cuestiones semánticas han dedicado enorme atención —como es lógico— al verbo «saber» [conocer]^[*]1. En general se distinguen tres acepciones: 1) «X sabe que p» implica que p es *verdad* (podemos denominarlo el elemento de *verdad*); 2) «X sabe que p» implica que X cree que p (los filósofos se han querellado a propósito del vocablo pertinente, algunos sostienen que debería ser «X *confía* en que p», o «X *está en posición de afirmar* que p»; a éste lo denominaremos el elemento de *confianza*); 3) «X sabe que p» implica que X posee evidencia de que p (creo que aquí la palabra «evidencia» es totalmente desacertada^[11], pero no importa para lo que aquí nos atañe; le denominaré el elemento *evidencia*!). Además, inherente al significado de la palabra «evidencia» es que nada puede literalmente ser evidencia de sí mismo: si X es evidencia de Y, X e Y han de ser cosas distintas.

A la vista de estos análisis, se han originado controversias acerca de si es pertinente decir (en casos como el que estamos considerando): «Jones sabe que le duele». Por un lado, los filósofos que adoptan la perspectiva del sentido común («Cuando me duele, sé que me duele») argumentan en esta línea: Sería totalmente falso decir que Jones *no* sabe que le duele; pero Jones tanto sabe como no sabe; por consiguiente, Jones sabe que le duele. En contra de estos filósofos podría argumentarse lo siguiente: «Jones no conoce a X» implica que Jones no está en posición de afirmar que X; por consiguiente, es totalmente erróneo decir «Jones no sabe que le duele». Pero este empleo de la Ley del Medio Excluido es un sofisma: en inglés, las palabras tienen *gamas de significados* y lo que se sostiene es que no es semánticamente correcto decir *ni* «Jones sabe que le duele» *ni* «Jones no sabe que le duele» (si bien la primera frase es, desde luego, menos engañosa que la segunda, ya que al menos cumple *una* de las condiciones del conocimiento, puesto que Jones está en posición de afirmar que le duele). (De hecho, están presentes los elementos de *verdad* y de *confianza*, y es el elemento evidencial el que agua la fiesta.)

No voy a discutir aquí esta cuestión^[12]; lo que ahora me interesa son más bien las similitudes entre ambas cuestiones. Por ejemplo, podemos decidir aceptar las dos afirmaciones (en tanto que «no desviadas», «lógicamente correctas», «no contradictorias en sí», etc.):

(a) La máquina verificó que estaba en el estado A,

(b) Jones sabía que le dolía,

o podemos rechazarlas. Si rechazamos (a) y (b), podemos encontrar otras formulaciones que sean semánticamente aceptables; por ejemplo, para (a) «La

máquina estaba en el estado A y eso hizo que escribiera “Estoy en el estado A”»; para (b) «A Jones le dolía y lo mostró diciendo “Me duele”».

Por el contrario, si se acepta (a) y (b), hay que hacer frente a las preguntas (a¹): «¿Cómo verificó la máquina que estaba en el estado A?»; y (b¹): «¿Cómo supo Jones que le dolía?»

Y, si se considera que estas preguntas tienen algún tipo de respuesta, serán respuestas degeneradas; por ejemplo: «Estando en el estado A» y «Al dolerle».

A estas alturas, está muy claro —creo yo— que la dificultad tiene en ambos casos el mismo origen. Es decir, la dificultad se origina en el hecho de que el «informe verbal» («Estoy en el estado A», o «Me duele») procede directamente del estado sobre el que se «informa» y no requiere «computación» o «evidencia» suplementarias para llegar a la «respuesta». Y las discordancias filosóficas sobre «cómo hablar», se refieren en el fondo al hallazgo de una terminología para la descripción general de los procesos cognoscitivos que no sea engañosa en este caso particular. [Obsérvese que la respuesta epistemológica tradicional a (b¹) —es decir, «por introspección»— es falsa en cuanto a los hechos de este caso, ya que implica claramente la ocurrencia de un acontecimiento mental (el «acto» de introspección) distinto al de la sensación del dolor].

Finalmente, supongamos que la máquina está equipada para «barrer» a una máquina próxima T¹. Entonces, vemos que la pregunta «¿Cómo verifica T que T¹ está en el estado A?» es una pregunta perfectamente lógica, tanto como «¿Cómo verifica T que el trimilésimo dígito de it es éste y aquél?» En ambos casos la respuesta presupone la descripción de todo un «programa» (aparte de explicar la *lógica* del programa, en caso necesario). Además, habría que decir algo sobre los contextos físicos que relacionan a T con T¹ (disposición de órganos sensoriales, etc.), y no limitarse a describir los estados internos de T, debido a que T, ahora, responde a una pregunta *empírica* y no a una matemática. Igualmente, «¿Cómo sabía Sherlock Holmes que a Jones le dolía?» es una pregunta perfectamente lógica de respuesta bastante complicada.

3. Estados «mentales» y estados «lógicos»

Consideremos las dos preguntas:

- 1) ¿Cómo sabe Jones que le duele?
- 2) ¿Cómo sabe Jones que tiene fiebre?

La primera, como vimos en el apartado anterior, es un tanto rara. La segunda bastante lógica. De hecho, si Jones dice «Me duele», nadie va a replicarle «Te equivocas» (se le *podría* replicar: «Has tenido un desliz de lenguaje» o «Mientes», pero no «Te equivocas»). Por el contrario, si Jones dice «Tengo fiebre», el médico que acaba de tomarle la temperatura puede, con suficiente lógica, responderle: «Te

equivocas». Y no es que el médico quiera decir que Jones ha cometido un error lingüístico, haya mentido o se confunda.

Podría pensarse, sin embargo, que, aunque la diferencia entre afirmaciones relativas al estado personal y afirmaciones relativas al estado de los demás presenta una analogía en el caso de las máquinas, la diferencia que acabamos de sugerir entre las afirmaciones sobre el estado «mental» personal y las afirmaciones sobre el estado «físico» personal, en el habla tradicional, no presenta analogía alguna. Pero no es así. Pasaremos a explicar cuál es la analogía.

En primer lugar, volvamos al concepto de máquina de Turing. Cuando se describe una máquina de Turing mediante una «tabla de instrucciones», se la describe como algo que consta en una cinta, un dispositivo de impresión, un dispositivo «traceador» (que puede simplemente ser un punto de la máquina que, en un determinado momento, se alinea con una casilla de la cinta) y una serie finita de «estados» (A, B, C, etc.). (A continuación, los denominaremos *estados lógicos* para diferenciarlos de otros estados de los que hablaremos en breve.) Aparte de esto, se la describe únicamente señalando las reglas deterministas que estipulan el orden en que se suceden los estados y lo que en ellos se escribe.

Concretamente, la «descripción lógica» de una máquina de Turing no incluye especificación alguna sobre la *naturaleza física* de esos «estados», ni, naturalmente, sobre la naturaleza física de la máquina entera. (¿Consta de relés electrónicos, cartón, empleados humanos sentados frente a escritorios, o qué?) En otras palabras, una «máquina de Turing» es una máquina *abstracta* susceptible de ser construida físicamente en un número casi infinito de modos distintos.

Sin embargo, en cuanto una máquina de Turing queda físicamente construida, sucede algo interesante. Aunque la máquina, desde el punto de vista lógico, sólo posee los estados A, B, C, etc., desde el punto de vista del ingeniero posee un número casi infinito de «estados» suplementarios (aunque no en el mismo sentido de «estado», los denominaremos *estados estructurales*). Por ejemplo, si la máquina consta de tubos de vacío, algo que puede suceder es que falle uno de esos tubos, lo cual sitúa a la máquina en un «estado» distinto desde el punto de vista físico, si no lógico. Si la máquina es de cartón y se opera manualmente, uno de sus posibles estados «no lógicos» o «estructurales» es que el cartón se doble. Y así sucesivamente.

Una máquina de Turing construida físicamente puede no disponer de medios para verificar su propio estado estructural, como tampoco un ser humano dispone de medios para verificar en un momento dado el estado de su apéndice. Sin embargo, es muy conveniente dotar a la máquina de «órganos sensoriales» electrónicos que la permitan explorarse y detectar averías menores. Pueden visualizarse estos «órganos sensoriales» haciendo que impriman ciertos símbolos en una «cinta de entrada» que la máquina «examina» de vez en cuando. (Una dificultad secundaria es que el «informe» de un órgano sensorial llegue a ocupar varias casillas de la cinta a la vez, aunque no tiene importancia, pues es bien sabido que el efecto de la «lectura» de un

número finito de casillas es posible mediante un programa que sólo requiera la exploración de las casillas una por una.)

(A título de digresión, señalaré que las primeras computadoras digitales que se construyeron no contaban con dispositivos de este tipo. Por el contrario, tenían, de hecho, más de 3.000 tubos de vacío, y algunos fallaban cuando menos se esperaba. Por eso se evidenció enseguida la necesidad de programas «rutinarios» de autocomprobación)^[13].

Una máquina capaz de detectar al menos parte de sus estados estructurales está en posición muy análoga a la de un ser humano capaz de detectar, con diversos grados de fiabilidad, algunos de los desajustes de su propio cuerpo. Supongamos, por ejemplo, que la máquina «escribe»: «Avería en el tubo de vacío 312». La pregunta «¿Cómo verifica la máquina que falla el tubo de vacío 312?» es una pregunta totalmente lógica, y en la respuesta puede integrarse una referencia a su estructura física («órganos sensoriales», etc.) y a la «estructura lógica» (programa para «leer» e «interpretar» la cinta de entrada).

Si la máquina escribe: «Avería en el tubo de vacío 312» cuando, en realidad, el tubo de vacío 312 funciona, el error puede ser debido a una computación equivocada (durante la «lectura» e «interpretación» de la cinta de entrada) o a una señal incorrecta de un órgano sensorial. Por el contrario, si la máquina escribe: «Estoy en el estado A» y sólo lo hace porque la tabla contiene la instrucción: *Escribe: «Estoy en el estado A cuando estés en el estado A»*, no puede plantearse la cuestión de computación errónea. Incluso si un accidente hace que el mecanismo de escritura señale: «Estoy en el estado A» cuando la máquina *no* se halla en dicho estado, no se habría producido una «computación equivocada» (sino únicamente, por así decirlo, un «desliz verbal»).

Es interesante señalar que, del mismo modo que hay dos descripciones posibles del comportamiento de una máquina de Turing—el plano estructural del ingeniero y la «tabla de instrucciones»—, también hay dos descripciones posibles de la psicología humana. El enfoque «conductista» (incluyendo en esta categoría las teorías que recurren a las «ideaciones hipotéticas» sin discriminar a las «ideaciones» extraídas de la fisiología) pretende aportar una descripción fisicalista^[14] completa de la conducta humana, en términos relacionados con la química y la física. Esto corresponde a la descripción del ingeniero o del físico de una máquina de Turing construida físicamente, pero también podría hallarse una descripción más abstracta de los procesos mentales humanos, en términos de «estados mentales» (prescindiendo de especificar la actuación física, si la hay) y de «impresiones» (que desempeñan el papel de los símbolos en la cinta de la máquina), descripción en la que se especificarían las leyes que rigen el orden en que se suceden estos estados y su relación con la verbalización (o, en cualquier caso, con el pensamiento verbalizado). Esta descripción, que sería la analogía de una «tabla de instrucciones», es en realidad el programa que debería aportarnos la psicología clásica. Está difundida la opinión de

que la psicología clásica ha fallado por motivos *metodológicos*; yo sugeriría, a la luz de la anterior analogía, que el fallo es imputable a motivos empíricos, ya que los estados mentales y las «impresiones» de los seres humanos no constituyen las «configuraciones» de una máquina de Turing.

Personalmente encuentro muy sugerente la analogía que hemos citado entre estados lógicos de una máquina de Turing y estados mentales de un ser humano, por un lado, y estados estructurales de una máquina de Turing y estados físicos de un ser humano, por otro. Creo que, en concreto, un estudio en profundidad de esta analogía nos permitiría progresar en el esclarecimiento del concepto de «estado mental» en cuestión. Este «estudio en profundidad» no se ha llevado a cabo, al menos por mi parte, pero me gustaría señalar, para quienes estén interesados, algunas características que parecen diferenciar los estados lógicos y mentales de los estructurales y físicos respectivamente:

- 1) la organización funcional (resolución de problemas, reflexión) del ser humano o de la máquina puede describirse en términos de las secuencias de estados mentales o lógicos respectivamente (y la simultánea verbalización), sin necesidad de referirse a la naturaleza de la «realización física» de estos estados;
- 2) los estados parecen íntimamente relacionados con la *verbalización*;
- 3) en el caso de pensamiento racional (o computación), el «programa» que determina qué estados suceden a otros, etc., está abierto a una crítica racional.

4. «Identidad» mente-cuerpo

El último plano en que hemos de comparar los seres humanos con las máquinas conlleva la problemática de *identificar* los estados mentales con los estados físicos correspondientes (o los estados lógicos con los estados estructurales correspondientes). Como señalábamos al principio, todos los argumentos a favor y en contra de esta identificación pueden discutirse perfectamente en términos de la máquina de Turing.

En los años 30, por ejemplo, Wittgenstein utilizaba el siguiente argumento: Si observo una imagen y, al mismo tiempo, observo mi estado cerebral (mediante un instrumento adecuado), observo *dos* cosas, no una. (Es de suponer que se trata de un razonamiento *contrario* a la identificación.) Pero podemos perfectamente imaginar una máquina de Turing, más «lista», que «razone» así: «Cuando escribo “Estoy en el estado A”, no tengo que emplear mis “órganos sensoriales”. Cuando empleo mis “órganos sensoriales” y comparo las ocasiones en que estoy en el estado A con aquéllas en que está actuando el *flip-flop* 36, estoy comparando *dos* cosas, no una». ¡Y no creo que el razonamiento de este Wittgenstein mecánico sea muy convincente!

Por el contrario, Russell llevó la tesis de la «identidad» al extremo absurdo de sostener que todo lo que *vemos* son porciones de nuestro cerebro. De forma análoga, un Russell mecánico podría «argumentar» que «todo lo que observo son mis tubos de

vacío». Los dos «Russell» están equivocados: el ser humano observa los acontecimientos del mundo exterior, y el proceso de «observación» presupone acontecimientos cerebrales, pero no por eso diremos necesariamente que, «en realidad», observa su cerebro.

También la máquina T «observa», por ejemplo, latas de sopa de tomate (si su función es clasificar latas de sopa), y el proceso de «observación» presupone el funcionamiento de tubos de vacío. Pero no por eso se dirá que la máquina observa «en realidad» sus propios tubos de vacío.

Pero consideremos argumentos más serios sobre el tema. Al principio señalé que el carácter *sintético* de la afirmación 1) «Me duele si, y sólo si, se estimulan mis fibras C», se había empleado como argumento a favor de la tesis de que las «propiedades» (o los acontecimientos), o los «estados») de «que se estimulen las fibras C» y de «sentir dolor» no pueden ser las mismas. Hay, cuando menos, dos motivos por los que se demuestra que no es un argumento sólido: (A) la distinción «analítico-sintética» no es tan clara, en particular en lo que respecta a las leyes científicas y (B) el criterio utilizado en este caso para identificar las «propiedades» (o «acontecimientos», o «estados») es muy cuestionable.

Con respecto al punto (A), ya he argumentado en otro trabajo^[15] que no puede clasificarse alegremente las leyes científicas fundamentales en «analíticas» y «sintéticas». Considérese, por ejemplo, la clase de sesgo conceptual que implica la transición de la geometría euclidiana a la no euclidiana, o el que implicaría el abandono de la ley de conservación de la energía. Es una distorsión decir que las leyes de la geometría euclidiana (mientras estuvieron vigentes) eran «analíticas» y que Einstein se limitó a «cambiar el sentido de las palabras». Al contrario, fue precisamente porque Einstein *no* cambió el sentido de las palabras —puesto que, en realidad, se refería a trayectorias más cortas del espacio en que vivimos, nos movemos y existimos— por lo que la teoría de la relatividad resultó tan incomprensible cuando se formuló por vez primera. ¡Decir que se regresaba al mismo punto al moverse uno en una sola dirección y en línea recta! Efectivamente, la adopción de la teoría de la relatividad supuso asumir un nuevo sistema de conceptos, pero eso no es decir «la adopción de un nuevo sistema de etiquetas verbales».

Pero, si distorsión es asimilar la revisión de las leyes científicas fundamentales a la adopción de nuevas convenciones lingüísticas, igualmente erróneo es seguir a los filósofos convencionales de la ciencia y asimilar el cambio conceptual que instauró Einstein con la clase de cambio que se produce si descubrimos un cisne negro (¡cuando previamente habíamos asumido que todos los cisnes eran blancos!) Las leyes fundamentales son como principios de matemática pura (como puso de relieve Quine), en el sentido de que no pueden desmoronarse por efecto de experimentos aislados, pues siempre podemos ajustarnos a las leyes y explicar los experimentos de diversos modos, más o menos *ad hoc*. Y —a pesar del regusto peyorativo de «*ad hoc*»— es hasta *racional* hacerlo en el caso de importantes teorías científicas, *en*

tanto no exista otra teoría aceptable. Por eso se tardó un siglo en la génesis del concepto —y no bastó un solo experimento— para desbancar la geometría euclidiana. De igual modo, tampoco podemos actualmente describir *ningún* experimento que por *sí solo* destruya la ley de conservación de la energía, a pesar de que esta ley no es «analítica» y podría abandonarse si un nuevo Einstein aportara para ello buenas razones *teóricas* y los experimentos pertinentes.

Como ha dicho Hanson^[16], nuestros conceptos están formados por teorías «ensambladas» y, por ello, abandonar una teoría científica fundamental sin presentar otra alternativa, equivaldría a «dejar que nuestros conceptos se derrumben». Por el contrario, aunque hayamos *podido* sostener que «todos los cisnes son blancos», ante la evidencia contraria, *no* vendría al caso, ya que los conceptos en cuestión no se *fundaban* en la aceptación de éste u otro principio rival, al igual que los conceptos geométricos se basan en la aceptación, no necesariamente de la geometría euclidiana, sino de *una* geometría.

No niego que *hoy día* cualquier «correlación» recién descubierta, al estilo de: «Uno se halla en el estado mental Ψ si, y sólo si, se halla en el estado cerebral Φ », sería *en principio* una *simple* correlación, una mera «generalización empírica», pero sostengo que el caso interesante es el que se produciría si hubiéramos elaborado un *sistema* teórico pormenorizado de tales «correlaciones». En tal caso, los comentarios científicos serían muy otros. Los científicos empezarían a decir: «Es imposible *en principio* hallarse en el estado mental Ψ ». Y podría perfectamente darse el caso de que la «imposibilidad en principio» no pasara de ser lo que Hanson, con toda justicia, denomina imposibilidad *conceptual*^[17]: los científicos no podrían *concebir* (salvo un nuevo Einstein) que alguien se hallara en el estado mental Ψ sin hallarse en el estado cerebral Φ . Y, en concreto, ningún experimento podría *por sí solo* derrumbar las leyes psicofísicas que hubieran adquirido tal arraigo^[18]. ¿Está claro que en este tipo de situación científica no fuera correcto decir que Ψ y Φ son el *mismo* estado?

Además, los criterios para identificar «acontecimientos» o «estados» o «propiedades» no están nada claros. Un ejemplo de ley con la clase de arraigo a que nos referimos es la siguiente: La luz pasa a través de un orificio si, y sólo si, la radiación electromagnética (de tal y cual longitud de onda) pasa a través del orificio.

Es bastante evidente que esta ley no es una afirmación «analítica»; sin embargo, sería expresarse en terminología perfectamente científica decir: (i) la luz que pasa a través de un orificio y (ii) la radiación electromagnética (de tal y cual longitud de onda), que pasa a través de un orificio, son dos descripciones del mismo acontecimiento. (Efectivamente, en «lenguaje ordinario», no sólo no es necesario que las descripciones del mismo acontecimiento sean equivalentes, sino que puede hablarse hasta de descripciones incompatibles de un mismo acontecimiento.)

Sin embargo, puede sostenerse que las *propiedades* (al contrario que los acontecimientos) no pueden expresarse con descripciones distintas no equivalentes. Efectivamente, Frege, Lewis y Carnap han *identificado* propiedades con

«significados» (por lo que, *por definición*, si dos expresiones tienen significado distinto, es que «significan» propiedades distintas). A mí me parece muy dudoso, pero supongamos que fuera correcto. ¿Cuáles serían las consecuencias? Habría que admitir que, por ejemplo, sentir dolor y tener estimuladas las fibras C son propiedades distintas. Pero, en el lenguaje de la «teoría constructiva» de la máquina de Turing utilizado al principio de este trabajo, habría también que admitir que «estar en el estado A» y «tener el *flip-flop* 36 en acción» son propiedades distintas. Efectivamente, está claro que las frases (i) «Estoy en el estado A» e (ii) «el *flip-flop* 36 actúa» no son sinónimas en el lenguaje de verificación de la máquina (tienen propiedades sintácticas distintas, y distintas «condiciones de expresión», es decir que la máquina tiene que emplear distintos «métodos de verificación»). Por lo tanto, quien desee argumentar sobre esta base respecto de la existencia del alma, tendrá que estar dispuesto a acoger en su regazo filosófico al alma de las máquinas de Turing.

5. Un argumento «lingüístico»

El último argumento que trataremos en relación con el tema de la identidad mente-cuerpo es un argumento «lingüístico» muy manido. Lo utilizó, por ejemplo, Max Black frente a Herbert Feigl en el seminario que inspiró este libro. Consideremos la frase:

1) El dolor *es idéntico* a la estimulación de las fibras C.

Es una frase desviada (es la línea de argumentación, aunque no con esta terminología): no hay una afirmación que pueda emplearse para establecer un contexto normal. Por lo tanto, si un filósofo lo propone como tesis, tiene que estar atribuyendo a las palabras un nuevo significado en lugar de expresar algún tipo de descubrimiento. Por ejemplo (argumentaba Max Black), se podría empezar a decir «Tengo fibras C estimuladas» en lugar de «Me duele», etc. Pero, en tal caso, se estaría dando *simplemente* a la expresión «tiene fibras C estimuladas» el nuevo significado de «le duele». La tesis es que, mientras las palabras mantengan su actual significado, la frase 1) es ininteligible.

Coincido en que la frase 1) es «desviada» en nuestro lenguaje actual. No concuerdo con que la frase 1) no pueda convertirse nunca en una frase normal, no desviada, si no cambia el significado actual de las palabras.

En resumen, el *quid* está en que lo que es «desviado» depende en gran medida del contexto, incluido nuestro nivel de conocimientos: y, con el desarrollo de nuevas teorías científicas, sucede constantemente que frases que antes no «se empleaban» y eran «desviadas» adquieren uso, no porque las palabras adquieran *nuevo* significado, sino porque el antiguo significado, fijado por el núcleo de uso común, *determina* un nuevo empleo dentro del nuevo contexto.

No hay nada que impida tratar de orientar la teoría lingüística hacia esa perspectiva, pero habría que contar con una teoría lingüística algo sofisticada para

ello. La verdadera cuestión no es de lingüística *sincrónica*, sino *diacrónica*^[19], no de «¿Es 1) *ahora* una frase desviada?», sino de «Si un cambio de conocimientos científicos (por ejemplo, el desarrollo de una red integrada de leyes psicofísicas de elevada "prioridad" en nuestra perspectiva del mundo, predominantemente científica) hiciera que 1) se convirtiese en una frase *no* desviada, ¿se habría producido necesariamente un cambio de significado en la palabra?» No es una cuestión sencilla.

Aunque no es el momento ni el marco adecuados para intentar elaborar una teoría semántica^[20], quisiera apuntar algo a propósito de esta cuestión.

En primer lugar, incluso si una frase que previamente era desviada empieza a emplearse de forma generalizada, no es necesario que haya cambiado el *significado* de ninguna palabra. Por lo tanto, la frase «Estoy a mil millas de distancia de ti», o su traducción al griego clásico, era sin lugar a dudas una frase desviada antes de la invención de la *escritura*, pero adquirió (no se le «dio», sino que *adquirió*) un uso normal con el invento de la escritura y la subsiguiente posibilidad de domicilio interpersonal distanciado.

Obsérvense los motivos por los que no decimos que palabra alguna (por ejemplo, «yo», «ti», «mil») de la frase haya cambiado de significado: (A) el nuevo empleo no era *arbitrario*, no era producto de *estipulación*, sino que representa una proyección^[21] automática de los usos comunes existentes en las palabras que forman la frase, dentro del nuevo contexto; (B) el significado de una frase es, en general, función de los significados particulares de las palabras que la forman (de hecho, este principio subyace a todo el concepto del significado verbal, y, así, si decimos que la *frase* ha cambiado de significado, tendremos que enfrentarnos a la pregunta «¿,Qué palabra ha cambiado de significado?» Pero esto sería, en este caso, decididamente una embarazosa pregunta.

En el caso que hemos expuesto, el nuevo contexto era producto de una nueva tecnología, pero un nuevo conocimiento teórico puede ejercer similar efecto en el lenguaje. (Por ejemplo, «dio toda la vuelta al mundo» sería una frase desviada en una civilización que ignorase que la Tierra es redonda.) Malcolm trató un caso de este tipo; Empezamos a disponer de los medios para determinar, en base a diversos indicadores fisiológicos (electroencefalogramas, movimientos oculares durante el sueño, alteraciones de la presión sanguínea, etc.), cuándo empiezan y acaban los sueños. Por lo tanto, la frase «Está en medio de un sueño» puede adquirir algún día uso generalizado. A este respecto el comentario de Malcolm consistía en que, en tal caso, se habría *dado* un uso a las palabras. Es evidente el error de Malcolm, creo yo; el caso en el que una frase adquiere un uso *debido* al significado de las palabras, está en las antípodas del caso en el que literalmente se *da* un uso a las palabras (es decir, aquél en el que se estipulan significados para las expresiones). El enunciado «realista» de este caso es, en mi opinión, totalmente correcto: la frase no tenía previamente un uso porque no podíamos decir cuándo empezaban y concluían los sueños. Actualmente empezamos a disponer de medios para determinarlo y, por lo

tanto, empezamos a hallar ocasiones en las que es natural emplear la frase. (Obsérvese que, en la definición de Malcolm, no se explica el hecho de que demos a esta frase este uso.)

Quizás alguien aduzca que no debe confundirse el cambio de significado con el cambio de distribución^[22] y que los progresos científicos y tecnológicos producen muchas veces en éstas modificaciones que en rigor no pueden considerarse cambios en aquél. Pero puede argumentarse que, aunque pueden preverse las circunstancias en las que la frase «Dio toda la vuelta al mundo» se convertiría en no desviada, no se pueden prefigurar las circunstancias en las que^[23] «estado mental es idéntico al estado cerebral Φ » se convierta en una frase no desviada.

Para ello es necesario hablar de una clase importante de «es», el «es» de *identificación teórica*. El uso del «es» en cuestión queda ilustrado en las siguientes frases:

2) La luz es radiación electromagnética (de tal y cual longitud de onda).

3) El agua es H₂O.

Lo que presupone la aceptación científica de, por ejemplo, 2) fue lo siguiente: antes de llegar a esta identificación, existían dos *corpus* teóricos distintos, la teoría óptica (cuya naturaleza ha descrito magistralmente Toulmin en su obra sobre filosofía de la ciencia) y la teoría electromagnética (representada por las ecuaciones de Maxwell). La decisión de *definir* la luz como «radiación electromagnética de tal y cual longitud de onda» se justificó científicamente mediante las siguientes consideraciones (como se ha señalado en numerosas ocasiones):

(a) Posibilitaba la *derivación* de las leyes ópticas (hasta una primera aproximación) de otras leyes físicas más «básicas». Y, por ello, aunque no se hubiera conseguido más que eso, esta identificación teórica habría supuesto un paso- hacia la simplificación de la estructura de las leyes científicas.

(b) Posibilitaba la derivación de *nuevas* predicciones en la «reducida» disciplina (es decir, la óptica). En concreto, hacía viable la predicción de que, en determinados casos, las leyes de la óptica geométrica *no* eran aplicables. (Cf. los famosos comentarios de Duhem sobre la reducción de las leyes de Kepler a las de Newton.)

Tratemos de imaginar las circunstancias en las que una identificación teórica de los estados mentales con los estados fisiológicos puede concordar con un procedimiento científico irreprochable. En términos generales, se requiere que no se expongan *meros* «correlatos» para estados subjetivos, sino algo mucho más elaborado; por ejemplo, que conozcamos estados físicos (microestados de los procesos centrales, pongamos por caso) en base a los que sea posible no simplemente *predecir* la conducta humana, sino dar de ella una explicación causal.

Para evitar «errores de categoría» hay que restringir con sumo cuidado el concepto de «explicar la conducta humana». Supongamos que una persona dice «Me encuentro mal». Su conducta, descrita en una serie de categorías, es: «afirmación de que se encuentra mal», y la explicación sería «Dice que se encuentra mal porque tiene

hambre y cefalea». No pretendo sugerir que el acontecimiento «Jones afirmando que se encuentra mal» pueda explicarse en términos de las leyes de la *física*, pero hay otro acontecimiento muy relevante, el de «El cuerpo de Jones produciendo ondas sonoras». Desde un punto de vista, es un «acontecimiento distinto» al de que Jones afirme que se siente mal, pero (valiéndonos de una observación de Hanson) no vendría al caso señalar que son dos acontecimientos distintos si no existiese un sentido en el que son *el mismo* acontecimiento. Y es el sentido en el que son el «mismo acontecimiento», y no el sentido en el que son «distinto acontecimiento». el que es relevante en este caso.

En resumidas cuentas: por «explicación causal de la conducta humana» entiendo explicar causalmente ciertos acontecimientos físicos (sensaciones corporales, producción de ondas sonoras, etc.) que, en el sentido a que aludíamos, son «lo mismo» que los acontecimientos que constituyen la conducta humana. Y no hay ningún «ryleismo» capaz de descartar lo que claramente es una probabilidad: que la ciencia física hará grandes progresos en este terreno.

Si estos grandes progresos se hicieran realidad, identificar teóricamente los «estados mentales» con sus «correlatos» presentaría las siguientes ventajas:

(1) Sería posible («hasta una primera aproximación», insisto) derivar de la teoría física las leyes clásicas (o generalizaciones pedestres) de la psicología «mentalista» de sentido común, tales como: «Las personas tienden a evitar las cosas con las que han tenido experiencias dolorosas».

(2) Sería posible predecir los casos (y son legión) en que falla la psicología «mentalista» de sentido común.

La ventaja (2) podría lograrse, claro está, sin necesidad de «identificación» (recurriendo a las leyes de correlación); pero la ventaja (2) podría también haberse logrado en el caso de la óptica, sin necesidad de identificación (asumiendo que la luz *acompaña* la radiación electromagnética, aunque no sea *idéntica* a ella). Pero el *doble* efecto de eliminar del todo ciertas leyes (en favor de definiciones teóricas), y aumentar el poder explicativo de la teoría, no habría podido lograrse de ninguna otra manera en ambos casos. La cuestión relevante es igualmente aplicable en el caso mente-cuerpo y en el caso luz-electromagnetismo. (Incluso el argumento del «lenguaje ordinario» podría haberse esgrimido en contra de la identificación de la luz con la radiación electromagnética).

[Dos cuestiones secundarias: (i) cuando llamo definición a «la luz es radiación electromagnética (de tal y cual longitud de onda)», no quiero decir que sea una afirmación «analítica». Pero realmente las *propiamente llamadas* definiciones en ciencia teórica *nunca* son virtualmente analíticas^[24]. (Quine señaló en cierta ocasión que se le ocurrían al menos nueve significados de «definición», ninguno de los cuales tenía nada que ver con la analiticidad). Claro que un filósofo podría objetar toda la *lógica* de la identificación teórica en base a que de nada sirve eliminar las «leyes» en favor de las «definiciones» si ambas son afirmaciones *sintéticas*. El hecho de que el

científico lo vea de modo totalmente distinto es otra muestra de lo inútil que es considerar la ciencia desde la perspectiva «¿analítica o sintética?» (ii) La aceptación de una identificación teórica, por ejemplo, «El dolor es la estimulación de las fibras C», no nos obliga a *intercambiar* los términos «dolor» y «estimulación de las fibras C» en el habla corriente, como sugirió Black. Por ejemplo, la identificación del «agua» con H_2O » está actualmente muy generalizada, pero a nadie se le ocurre decir «Tráeme un vaso de H_2O », si no es en broma].

Creo que con la exposición que antecede se logra (a) explicar el hecho de que frases como «El estado mental Ψ es idéntico al estado cerebral Φ » son desviadas en el lenguaje actual, y a la vez (b) dejar claro que las mismas frases pueden convertirse en *no* desviadas mediante un aumento adecuado de nuestra profundización científica en la naturaleza física y en las causas de la conducta humana. Las frases en cuestión no pueden emplearse actualmente para expresar una identificación teórica, por el hecho de haber establecido tal identificación. El acto de la identificación teórica no es un acto que pueda realizarse «a voluntad»; existen *condiciones previas* al mismo, como sucede con muchos actos, y actualmente no se satisfacen dichas condiciones. Por el contrario, si el tipo de teoría científica descrita anteriormente se materializase, las condiciones previas de la identificación teórica se cumplirían, del mismo modo que se cumplieron en el caso de la luz y el electromagnetismo, y las frases de la modalidad en cuestión requerirían *automáticamente* un uso, es decir la expresión de las identificaciones teóricas correspondientes. Insisto en que lo que diferencia el uso distinto *adquirido* del uso *dado* (y del «cambio de significado» propiamente dicho) es que el «nuevo uso» es una *proyección* automática a partir de usos existentes y no presupone estipulación arbitraria (salvo que haya algún elemento de «estipulación» en la aceptación de *cualquier* hipótesis científica, incluida «La Tierra es redonda»).

Hasta ahora hemos considerado únicamente frases del tipo^[25] «el estado mental Ψ es idéntico al estado cerebral Φ ». Pero ¿qué decir de la frase 3): ¿Son los estados mentales microestados cerebrales?

Esta frase, por así decirlo, no «da» identificación teórica *concreta* alguna; sólo dice que son posibles identificaciones teóricas inconcretas. Es el tipo de afirmación que haría Feigl. Y Black^[26] podría replicar que, al enunciar 3), Feigl había enunciado una extraña serie de palabras (es decir, una frase desviada). Es posible que Black tenga razón. Quizás 3) es desviada en el lenguaje actual, pero también es posible que nuestros descendientes, dentro de doscientos o trescientos años, consideren que Feigl expresaba algo pleno de sentido y que las objeciones lingüísticas a 3) eran una tontería. Y puede que también tengan razón.

6. Lingüística de máquinas

Consideremos la cuestión lingüística que acabamos de estudiar desde el punto de vista de la analogía entre el hombre y la máquina de Turing de que hablamos en este trabajo. Se verá que esta máquina de Turing sería probablemente incapaz, si no posee los «órganos sensoriales» adecuados, de elaborar una teoría correcta sobre su propia constitución. Por otro lado, «Estoy en el estado A» es una frase con un patrón concreto de ocurrencia en el «lenguaje» de la máquina. Si el «lenguaje» de ésta es lo bastante complejo, podría analizarse sintácticamente en términos de una serie finita de bloques constructivos básicos (morfemas) y reglas para la composición de una serie potencialmente infinita de «frases». En particular, sería posible distinguir las/rases *gramaticales*^[27] de las *no gramaticales* en el lenguaje de la máquina. También sería posible asociar las regularidades en la ocurrencia de frases (o, «describir usos de frases», según la jerga de Oxford), y atribuir «significados» a la serie finita de morfemas y a la serie finita de modalidades de composición, de tal manera que los «usos» de las diversas frases pudieran proyectarse eficazmente a partir de los significados de los morfemas y modalidades de composición aislados. En tal caso, podría distinguirse no sólo las frases «gramaticales» y «no gramaticales» en el «lenguaje de la máquina», sino también las «desviadas» y las «no desviadas».

Chisholm insistiría en que es impropio hablar de máquinas que emplean un lenguaje, y estoy de acuerdo. Es el motivo por el que, a veces, pongo entre comillas las palabras «lenguaje» y «significado», etc., para poner de relieve, cuando es necesario, que se emplean en sentido lato. Por otra parte, es importante admitir que las actuaciones de la máquina son totalmente *análogas* al lenguaje, tanto, que puede aplicárseles toda la teoría lingüística. Si el lector desea ampliar el tema, le recomiendo consulte a fondo una obra como *Estructuras sintácticas* de Chomsky y observe que *en ningún caso se da por supuesto que el Corpus de expresiones estudiadas por los lingüistas está producido por un organismo consciente*. A continuación puede consultar una obra de vanguardia sobre semántica empírica como *Análisis semántico* de Ziff y comprobar que la misma consideración es válida para la teoría *semántica*.

Dos observaciones más a este respecto: (i) como sostengo que la problemática mente-cuerpo es *estrictamente análoga* a la de la relación entre estados estructurales y lógicos, y no que los dos problemas sean *idénticos*, me basta con una *analogía* adecuada, en este caso, entre «lenguaje» de máquina y lenguaje humano; (ii) puede que Chisholm sostenga la imposibilidad de una semántica «conductista» del tipo considerado por Ziff (es decir, una semántica que no admita la «intencionalidad» como concepto primordial). Pero, aunque esto fuera cierto, no sería relevante, ya que, si *cualquier* teoría semántica es aplicable al lenguaje humano, habría que demostrar por qué una teoría totalmente *análoga* no sería aplicable al lenguaje de una máquina adecuada. Por ejemplo, si la «intencionalidad» desempeña una función como concepto primordial en la explicación *científica* del lenguaje humano, una elaboración teórica de similares relaciones *formales* con los «observables»

correspondientes tendría *igual* poder explicativo en el caso del «lenguaje» de máquinas.

Por supuesto que la objeción a la lingüística conductista sería, *en realidad*, una objeción en todos los sentidos a la lingüística científica. Pero considero justificado el descartar esta posibilidad.

Supongamos ahora que dotamos la «teoría constructiva» de la máquina de Turing de «órganos sensoriales», de forma que pueda obtener los datos empíricos necesarios para elaborar una teoría sobre su propia naturaleza.

En este caso, podría incorporar a su «lenguaje teórico» frases sustantivas susceptibles de «traducirse» por la expresión en lenguaje real de «*flip-flop* 36», y frases que podrían traducirse por «Actúa el *flip-flop* 36». Estas expresiones tendrían un significado y un uso muy distintos al del sentido y el uso de «Estoy en el estado A» del lenguaje de la máquina.

Si algún razonamiento «lingüístico» demuestra realmente que la frase «El dolor es idéntico a la estimulación de fibras C» es desviada, el mismo razonamiento debe demostrar que «El estado A es idéntico a la actuación del *flip-flop* 36» es desviada en el lenguaje de la máquina. Si hay un razonamiento que demuestre que «el dolor es idéntico a la estimulación de las fibras C» no puede convertirse en una frase no desviada (considerando, en este caso, *diacrónicamente* el idioma) si previamente no cambian de significado las palabras, el mismo argumento aplicado a «la lingüística diacrónica del lenguaje de la máquina» demostraría que la frase «El estado A es idéntico a la actuación del *flip-flop* 36» no podría convertirse en no desviada en el lenguaje de la máquina, a menos que las palabras cambiasen previamente de significado. En resumen: cualquier argumento filosófico que se haya empleado en relación al problema mente-cuerpo, desde el más antiguo e ingenuo (por ejemplo, se ve claramente que «los estados de inconsciencia son distintos a los estados físicos») hasta el más sofisticado, tiene su exacta contrapartida en el caso del «problema» de los estados lógicos y los estados estructurales de las máquinas de Turing.

7. Conclusión

Creo que la moraleja está bien clara: ya no se puede creer que la problemática mente-cuerpo sea una cuestión teórica genuina, o que la «solución» arroje el mínimo rayo de luz en el mundo en que vivimos, porque está bastante claro que ningún adulto en pleno uso de sus facultades toma en serio el problema de la «identidad» o «no identidad» de los estados lógicos y estructurales de una máquina; no porque la respuesta sea obvia, sino porque es evidente que es poco importante *cuál* sea la respuesta. Pero, si el denominado «problema mente-cuerpo» no es más que una concepción distinta de las mismas perspectivas lógicas y lingüísticas, entonces, la cuestión resulta vacua y verbal.

Muchas veces constituye un enfoque importante la consideración de que dos problemas de distinta temática son el mismo en todos sus aspectos lógicos y metodológicos. En tal caso, el enfoque discurre sobre la aceptación de que cualquier conclusión a que se llegue en el caso de la problemática mente-cuerpo tiene que extraerse, y *por idénticos motivos*, en el caso de la máquina de Turing. Pero está claro (como es obvio) que, por ejemplo, la conclusión de que los estados lógicos de las máquinas de Turing son irremediabilmente distintos de sus estados estructurales, aun siendo correcta, sólo representaría un hallazgo puramente *verbal*, y entonces la misma conclusión, *alcanzada con el mismo argumento* en el caso humano, tendría que representar un mero hallazgo verbal. Por decirlo de otro modo, si se identifica el problema mente-cuerpo con cualquier otro problema que rebase un interés estrictamente conceptual (por ejemplo, la cuestión de si los seres humanos tienen o no «alma»), necesariamente, o *bien* (a) ningún argumento empleado por los filósofos arroja *la mínima* luz al respecto (y ello independientemente de la tendencia del argumento), o (b) algún argumento filosófico del mecanicismo es correcto, o (c) algún argumento dualista demuestra que *¡tanto* los seres humanos *como* las máquinas de Turing tienen alma! Dejo al lector la opción de juzgar cuál de las tres alternativas es plausible.

El sentir de los robots^[*]

Paul Ziff

¿Puede sentir un robot? Algunos dicen ¡claro!^[1], otros ¡claro que no!^[2]

1. Me refiero al tipo correcto de robots. Deben ser autómatas y, sin duda alguna, máquinas.

Supondré que fundamentalmente son máquinas computadoras con todos los microelementos y micro-mecanismos necesarios para el funcionamiento de esos prodigios de ingeniería. Además, supondré que los accionan microbaterías solares y que, en lugar de comer, se alimentan de luz.

Una vez sentado que estos robots son, sin duda alguna, máquinas, en todos los demás aspectos serán tan humanos como se quiera. Son de tamaño humano. Un vez vestidos y enmascarados son prácticamente indiferenciables de los seres humanos en casi todos los aspectos: su apariencia, sus movimientos, lo que dicen, etc. Por lo tanto, de no ser por la máscara, cualquier persona los tomaría por seres normales. Al no inspirar sospechas de que son robots, no infunden desconfianza. Pero, sin la máscara, todo su entramado metálico queda a la vista. De lo que aquí se trata no es de si podemos diluir la línea entre el hombre y la máquina y atribuir sentimientos a la máquina. La cuestión estriba en si podemos atribuir sentimientos a la máquina para diluir esa línea entre el hombre y la máquina.

2. ¿Pueden sentir los robots? ¿Pueden, por ejemplo, sentir cansancio o aburrimiento?

Ex hypothesi, los robots son mecanismos, no organismos; no son seres vivos. Puede haber un robot estropeado, pero no muerto. Sólo los seres vivos pueden sentir realmente.

Si digo «Se siente cansada», suele deducirse que de lo que se trata (o se trataba o se tratará, en el caso de hablar de espíritu)^[3] es de un ser vivo. De modo i más general, en el entorno lingüístico «...sentirse cansado» suele referirse únicamente a expresiones relativas a seres vivos. Supongamos que ustedes dicen: «El robot se siente cansado». La expresión «el robot» se refiere a un mecanismo. Por consiguiente, puede deducirse que no se trata de un ser vivo. Pero, por el enunciado de la expresión predicativa «...sentirse cansado», puede deducirse que se trata de un ser vivo. Por lo tanto, si se habla al pie de la letra y se dice: «El robot se siente cansado», se expresa una contradicción. En consecuencia, no se puede enunciar literalmente el predicado «...sentirse cansado» para «el robot».

Insisto en que ningún robot hará nunca lo que una persona. Y ello independientemente de cómo se construyan, o por complejos y variados que sean sus movimientos y operaciones. Los robots pueden calcular, pero no razonan en sentido estricto. Quizás agarren cosas, pero no las elegirían; matarán, pero no cometerían homicidio; presentarán excusas, pero

no las sentirían. Todo esto son actos que sólo las personas pueden realizar: *ex hypothesi*, los robots no son personas.

3. «Un robot muerto» es una metáfora, pero una «batería muerta» es una metáfora muerta: si se tratase de un robot, habría matado a su metáfora.

Lo que no quiero presuponer no tengo por qué presuponerlo. Una presunción puede paliarse; el sentido de una palabra puede acentuarse o restringirse, o cambiarse. Si uno desea que le entiendan, no hay que darle muchas vueltas: así de fácil. Señalando un cuadro entre otros muchos, digo: «¡Anda! *esto* es un *cuadro*». ¿Quiero acaso decir con ello que los otros no lo son? Claro que no. Sin embargo, el énfasis en *esto* es contrastante. Asimismo, si digo: «El robot, ese mecanismo, desde luego máquina y no ser humano, se siente cansado», no puede deducirse que esté hablando de un ser vivo.

Si digo de una persona; «Se siente cansado», ¿se piensa que estoy diciendo que es un ser vivo y nada más que eso? Si digo: «El robot se siente cansado», el predicado «...se siente cansado» significa lo que siempre significa, salvo que no se puede deducir que se trata de un ser vivo. Esta es la única diferencia.

Y lo que hemos dicho a propósito de «El robot se siente cansado» podría decirse perfectamente a propósito de «El robot es consciente», «El robot cogió mi gato» y así sucesivamente.

4. ¿Pueden sentir cansancio los robots? ¿Puede sentir cansancio una piedra? ¿Puede sentir cansancio el número 17? Está claro que no existen razones para pensar que el 17 se sienta cansado. Pero esto nada demuestra. Una persona puede sentir cansancio y puede que no haya nada —ni tiene por qué haberlo— que lo demuestre. Asimismo en el caso de un robot, una piedra o el número 17.

Aun así, el número 17 no puede sentir cansancio. Y digo esto, no porque, o no simplemente porque, no existan motivos para suponer que el 17 se sienta cansado, sino porque existen buenos motivos para no suponer que el 17 se sienta cansado y buenos motivos para no suponer que el 17 haya sentido alguna vez algo. En consecuencia, hay que considerar si existen motivos para suponer que los robots sientan cansancio y si hay buenos motivos para no suponer que los robots hayan sentido jamás algo.

5. Conociendo a George y viendo su aspecto, digo que se siente cansado. Conociendo a Josef y viendo su aspecto, no digo que se siente cansado. Pero, si tú no conoces a ninguno de los dos, entonces a ti pueden parecerte iguales.

También en cierto sentido, pero no en otro, pueden parecerme a mí iguales. Ahora bien, George, y no Josef, es el que parecerá cansado. Si me piden que indique la

diferencia puede que no exista nada relevante, no tiene por qué haber nada relevante que señalar, ya que la diferencia relevante puede ser la que existe entre la contemplación de un cuadro sin marco y la de un cuadro con marco. Sólo que, en este caso, hace de marco lo que sé respecto de ellos: y ustedes no pueden ver lo que yo sé de ellos.

(Hablando con los robots, puede decirse que el aspecto que para mí presentan las cosas, mi actual *output*, no será el mismo que para ustedes: el aspecto que para ustedes presentan las cosas, a pesar de que realmente ambos recibamos el mismo *input*, los mismos estímulos, y ello se debe a que sus mecanismos —el de ustedes— no estaban en el mismo estado inicial que el mío, debiéndose esto a una diferencia estructural o a una diferencial de entradas previas —*inputs*).

Si decimos de una persona que se siente cansada, generalmente lo hacemos no sólo basándonos en lo que vemos en esa ocasión, sino basándonos en lo que hemos visto en otras ocasiones y en base a cómo lo que hemos visto en otras ocasiones se relaciona con lo que vemos en esa ocasión concreta. Y esto es sólo para decir que, al determinar si una persona se siente o no cansada, intervienen las consideraciones observacionales y teóricas y, como en todos los casos, están inextricablemente relacionadas.

6. Supongamos que ustedes vienen a visitar conmigo a un actor en su casa. Está ensayando el papel de una persona abatida por el dolor. No advierte nuestra presencia, como podría suceder con una persona abatida por el dolor, y su actuación es impecable. Yo sé que es un actor y que está ensayando un papel, pero ustedes no. Ustedes preguntan: «¿Por qué está tan apenado?» y yo contesto: «No lo está». «Claro que sí», replican, «está muy afligido. ¡Mírele! ¡Demuéstrenos que no!», y, naturalmente, en este caso, no hay nada que demostrar.

Por eso planteó Turing^[4] la cuestión de si los autómatas podían pensar, ser conscientes, sentir, etc., del modo ingenuo siguiente: ¿qué tests no superaría un autómata? MacKay^[5] ha señalado que cualquier test sobre atributos mentales o de otra clase, a superar con arreglo a la actividad observable en un ser humano, puede superarlo un autómata. Por lo que uno se ve impulsado a decir qué habría de malo en la actuación de un robot.

No tiene que haber nada de malo en la actuación de un actor ni en la de un robot. Lo que no vale es que son actuaciones.

7. Supongamos que K es un robot. Una persona cualquiera ve a K y, sin saber que es un robot, dice: «K está cansado». Si le pregunto qué le hace pensarlo, me contestaría: «K ha estado todo el día cavando zanjas. Para qué decirle más, mire a K, si no tiene aspecto de cansado, que me aspen».

Por lo tanto, K parece cansado para una persona cualquiera. Eso no demuestra nada. Si yo sé que K es un robot, a mí K no me parecerá cansado. No se trata de lo que veo, sino de lo que sé. O no es lo que veo en esa ocasión, sino lo que he visto en otras. ¿Dónde? En un laboratorio de psicología de robots.

8. Si digo: «El robot se siente cansado»; el predicado «...se siente cansado» significa lo que suele significar, salvo que no puede deducirse que se trata de un ser vivo. Esa es la única diferencia.

Hablar de algo vivo e hablar de un organismo en un entorno. El entorno es el medio en que se da el comportamiento del organismo. La muerte es la extinción de la relación entre un organismo y su entorno. En la muerte me pluralizo y paso de ser uno a ser muchos. Me convierto en mis restos, me fundo con el medio.

Si pensamos en que los robots son montajes, hemos de pensar que se los desmonta. Es lo que hacemos nosotros en el laboratorio; cambiamos e intercambiamos sus partes, cambiamos e intercambiamos sus programas, los ponemos en marcha y los paramos, a veces en un estado, a veces en el otro, les quitamos la memoria, les hacemos recordar cosas que aún no habían sucedido, etc.

Y lo que comprobamos en el laboratorio es que no puede decirse razonablemente que un robot sienta. ¿Por qué no?

9. Porque no existe autenticidad psicológica en los robots, sino sólo en los constructores humanos de robots, porque el modo en que actúan los robots (dentro de un contexto concreto) depende fundamentalmente de cómo se les programe para actuar, porque podemos programar un robot para que actúe como deseamos, porque podemos programar un robot para que, al levantar una pluma, actúe como una persona cansada y, por el contrario, no muestre cansancio cuando levanta una tonelada, porque un robot no puede tener intencionalidad en lo que dice, porque podemos construir un robot que diga lo que queramos, porque desear el robot del prójimo sería como desear su coche y no a su mujer, porque los robots son sustituibles, porque los robots no tienen individualidad, porque pueden duplicarse sus partes consiguiendo dos máquinas virtualmente idénticas, porque pueden intercambiarse todas sus partes y seguirán siendo las mismas máquinas, porque pueden intercambiarse los programas de dos máquinas que tengan la misma estructura, porque...

Porque ningún robot actúa cansado, porque un robot sólo puede actuar como un robot programado para actuar como una persona cansada, porque supongamos que se programa a unos robots para que actúen como una persona cansada después de levantar una pluma, mientras que otros se programan de modo tal que nunca actúan como personas cansadas. ¿Acaso diremos: «¡Qué extraño!, algunos robots se sienten cansados casi enseguida mientras que otros nada»? O, supongamos que algunos se programan para que actúen como una persona cansada después de levantar algo azul y no algo verde. ¿Diremos acaso: «Algunos robots se sienten cansados cuando levantan cosas azules, pero no cuando levantan cosas verdes?», y ¿concluiremos que «algunos robots encuentran las cosas azules más pesadas que las cosas verdes»? El trabajo excesivo produce cansancio en las personas: ¿qué hará que un robot actúe como una persona cansada? Quizás el trabajo excesivo, o el trabajo leve, o cualquier cosa, ya que depende del capricho de la persona que lo construya (aunque este

capricho pueda modificarse a causa de cualquier anomalía que se manifieste en la red electrónica del robot y se produzcan las secuelas indeseables imprevistas de un programa mal concebido). ¿Diremos: «No viene a cuento plantearse qué es lo que hace que un robot se sienta cansado»? Y, si un robot actúa como una persona cansada, ¿qué? Algunos robot pueden estar programados para que requieran descanso, otros para que requieran más trabajo. ¿Diremos: «Este robot se siente cansado, ¡que vuelva al trabajo!»?

¿Qué sucedería si todo esto e hiciera con y a los seres humanos? ¿Qué sucedería si algún día desapareciera la diferencia entre el hombre y su entorno? Quizás ese día nos despertaríamos y veríamos que somos robots, pero no nos despertaríamos en un paraíso mecánico ni en un infierno automático, porque entonces no tendría sentido hablar de seres humanos con sensaciones, como también hoy día carece de sentido decir que los robots sienten.

Un robot actuaría como un robot.

El profesor Ziff y los robots^[*]

J. J. C. Smart

El profesor Ziff («El sentir de los robots», artículo anterior a éste) arguye que los robots no pueden sentir. Sólo los seres vivos sienten —dice—, y los robots no pueden ser seres vivos. Tanto la premisa como la conclusión me parecen controvertibles, aunque, en esta breve exposición, sólo pueda esbozar algunas de las razones en que me baso.

A) El concepto de «ser vivo», en tanto que opuesto a «robot», es impreciso.

(a) Vamos a suponer que la historia del *Génesis* es literalmente cierta. Adán y Eva serían robots. Serían artefactos creados por Dios. Si nos permitimos un sincretismo de la antigua teología con la biología moderna, podríamos incluso decir que Dios dio a Adán y Eva «programas», es decir, su conjunto de genes, probablemente moléculas de DNA, cuya función fuera registrar la información hereditaria.

(b) Consideremos el mecanismo autorreproductor de von Neumann (John von Neumann, «The General and Logical Theory of Autómata», *Cerebral Mechanisms in Behaviour*, págs. 1-31, Seminario Hixon, 1951). En este sentido, ¿serían por ello los descendientes de tal mecanismo menos seres vivos que los descendientes de Adán y Eva? Incluso podemos suponer pequeñas alteraciones aleatorias en la parte orgánica en que está impreso su diseño. Estas máquinas podrían evolucionar por selección natural, de arrollando tendencias y capacidades distintas a las de la máquina original.

Juntando (a) y (b) e demuestra lo imprecisa que es la diferenciación entre un artefacto (lo bastante complejo) y un ser vivo. Personalmente no me parece nada sorprendente, y me inclino a aceptar la tesis fisicalista de que los seres vivos son mecanismos fisio-químicos muy complicados.

(B) No veo por qué «esto siente» equivale a «esto es un ser vivo», si en «esto es un ser vivo» se considera que existe la implicación de «esto no es un artefacto». Ninguno de los artefactos que cualquiera de no otro hayamos conocido en la práctica son lo bastante complejos como para permitirse la afirmación de que sientan. Por lo tanto, actualmente no tiene por qué existir problema alguno si deducimos «esto es un ser vivo» cuando nos dicen «esto siente». Sin embargo, no tiene por qué ser necesariamente una implicación lógica. Quizás en el futuro veamos ejemplos contrarios.

(C) Supongamos que construimos un robot de tal complejidad que fuera capaz de aprender nuevos actos y aptitudes igual que lo hace un niño. (Compárese esto con la

máquina infantil de Turing, en «Maquinaria de computación e inteligencia»). Podría hasta convertirse en filósofo, dar conferencias y pronunciar discursos como cualquier ser humano. ¿Por qué no íbamos a afirmar entonces que lo que dijera tuviese intencionalidad? No se parecería en nada a la máquina de Ziff con disco fonográfico interno.

En pocas palabras, por todo esto, encuentro poco convincentes los argumentos de Ziff. Me temo que puedo haber malinterpretado su significado, pero pienso que Ziff podría haber aclarado sus intenciones.

Robots, S. A.^[*]

Ninian Smart

Paul Ziff, hablando de los robots (artículo pág. 151), enumera una serie de razones por las que no debe decirse que se sienten cansados, etc. Me apresuro a señalar que sus razones no persuadirían a un determinista, salvo en cuanto que demuestra este punto: que los robots no sienten porque no tienen cuerpo propiamente dicho. Mi argumentación es que esto no contribuye gran cosa a demostrar por qué ciertos artefactos no sienten. Por supuesto, uno no tiene por qué ser determinista; pero, si no se es, y a menos que de forma irrealista vaya uno a compartimentar a los seres humanos, las razones por las que éstos se sienten cansados, y los robots no, deben profundizarse mucho más. No voy a discutir todas las razones que esgrime Ziff, me centraré únicamente en algunas de las que expone en el apartado 9 de su artículo, por considerarlas cruciales. Ziff argumenta que los robots no sienten porque:

1. «El modo en que actúa un robot en un contexto concreto depende fundamentalmente de cómo esté programado^[1] para actuar». Por mor de simplicidad introduzco el concepto de Naturaleza en representación de la suma de causas que concurren en la creación de un ser humano considerado desde el primer momento de la concepción o en cualquier fase ulterior de su vida. ¿Qué hay de malo en que el determinista diga que el modo en que actúa una persona, en un contexto concreto, depende fundamentalmente de cómo la Naturaleza le programe para actuar? Programas sutiles, si se quiere, mucho más sutiles que los programas de las computadoras, pero los sutiles circuitos celulares determinan de todos modos la manera en que actuó, en una determinada situación.

2. «Podemos programar un robot para que actúe como deseemos». La Naturaleza no quiere las cosas, por lo que, en cierto modo, no hay paralelismo posible. Pero existe una alternativa, ya que la entidad de un efecto no tiene por qué resultar afectada por la entidad de la causa. ¿Expone Ziff nuestro sentido de dominio sobre los robots? No creo, porque podríamos adquirir un dominio aparentemente igual sobre el reino animal por medio de una cría depurada (que diera un gato que no atacase a los ratones, aunque no por ello tendríamos un gato que no siente).

No, la particular clase de dominio procede de lo siguiente:

3. «Porque un robot puede programarse para que actúe como una persona cansada al levantar una pluma y no cuando levanta una tonelada.» El dominio radica en que la misma estructura puede programarse de formas opuestas (quizá sea imposible la cría

de gatos que no ataquen a los ratones, si no cambiamos su programa). Para el determinista, el problema radica en que el robot de Ziff no posee un cuerpo propio: no está plenamente constituido. La Naturaleza dota al gato de su programa (o serie de programas) con su cuerpo y, para un nuevo programa, se requiere un nuevo cuerpo.

4. «Se pueden duplicar todas sus partes.» La Naturaleza sustituye de vez en cuando todas o casi todas las partes. Y la Naturaleza produce mellizos idénticos. En consecuencia, la Naturaleza podría producir ventillizos, y es un mundo posible, aunque triste, aquél en el que los seres humanos salgan del vientre materno con la regularidad y la similitud de los Cadillacs.

Estos ejemplos parecen demostrar, para el determinista que lea a Ziff, que es el cuerpo lo que cuenta. Ziff manifiesta al principio de su artículo que se refiere al tipo correcto de robots, uno que sea en esencia una máquina computadora^[2], y lo contrasta con un organismo vivo que está aislado de su entorno y en el que la muerte es la fusión con el entorno^[3]. En pocas palabras: para él, robot es algo con las mismas válvulas, distintos programas y organismo, algo con el mismo cuerpo y el mismo programa (o serie de programas).

Luego, la corporeidad se reduce esencialmente a esto: un robot toma cuerpo si, y sólo si, para volver a programarlo, es necesario dismantelarlo y reconstruirlo de modo distinto. Basta para apuntalar el punto crucial número 3 de la argumentación Ziff. Pero, si Ziff puede concebir un robot capaz de hacer los movimientos de cansancio, también se puede concebir un espécimen más eficaz que no puede programarse para que actúe cansado al levantar una pluma. ¿Pero deseamos decir que éste se siente cansado a veces?

Entonces, ¿cuándo vamos a delimitar la diferencia entre los robots con corporeidad que no sienten y los seres con corporeidad que sienten? La única sugerencia que se nos ocurre es que el dilema tiene algo que ver con la sutilidad y la complejidad. Ciertos juguetes mecánicos tienen corporeidad en este sentido, pero son toscos. Mas los criterios de sutilidad y complejidad son también demasiado vago para ser válidos. Y, en cualquier caso, alimentamos un regusto oculto por los fantasma. Un robot de Ziff no siente porque es una máquina dentro de un fantasma, no porque haya un fantasma dentro de la máquina. Pero decir que los robots no sienten porque no tienen cuerpo, aunque invita a reflexión, también nos deja insatisfechos.

Bibliografía

Se incluye todo el aparato bibliográfico que acompañaba los artículos recopilados en este libro (junto con el lugar y fecha de la edición original). Se citan también otros artículos de interés, aunque no hemos pretendido reunir una lista exhaustiva de libros y artículos sobre el tema «Mentes y máquinas», debido en parte a limitaciones de espacio, pero principalmente, porque ya existen excelentes bibliografías: Armer, Minsky (en particular su *Bibliography to the Literature on Artificial Intelligence*), o Simmons and Simmons.

- Anisimov, S. F., *Man and Machine (Philosophical Problems of Cybernetics)*, Report No. JPRS 1990-N. Washington, D.C.: Joint Publications Research Service, 1959.
- Armer, Paul, *Attitudes Toward Intelligent Machines*. Santa Monica, cal.: The RAND Corporation, 1962.
- Ashby, W. R., «Can a Mechanical Chess Player Outplay Its Designer?» *British Journal for the Philosophy of Science*, Vol. III (1952), 44-47.
- , *Design for a Brain*. Nueva York: John Wiley & Sons, Inc., 1954.
- Blake, D. V. and A. M. Uttley, eds., *Proceedings of a Symposium on Mechanization of Thought Processes*, 2 vols. Londres: Her Majesty's Stationery Office, 1959.
- Burks, Arthur W., «Computation, Behavior, and Structure in Fixed and Growing Automata», *Behavioral Science*, Vol. VI, No. 1 (1961), 5-22.
- Butler, Samuel, *Erewhon*. Caps. 23-25.
- Carnap, Rudolf, «The Interpretation of Physics», en *Readings in the Philosophy of Science*, H. Feigl y M. Scriven, eds., pp. 309-318. Nueva York: Appleton-Century-Crofts, Inc., 1953.
- , «The Methodological Character of Theoretical Concepts», en *Foundations of Science and the Concepts of Psychology and Psychoanalysis*, H. Feigl y M. Scriven, eds., Minnesota Studies in the Philosophy of Science, Vol. I, pp. 38-76. Minneapolis: University of Minnesota Press, 1956.
- Chappell, V. C., ed., *The Philosophy of Mind*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962.
- Chomsky, Noam, *Syntactic Structures*. The Hague: Mouton & Co., 1957.
- Church, Alonzo, *Introduction to Mathematical Logic*. Princeton: Princeton University Press, 1956.

- , «An Unsolvable Problem of Elementary Number Theory», *American Journal of Mathematics*, Vol. LVIII (1936), 345-363.
- Cohen, Jonathan, «Can There Be Artificial Minds?» *Analysis*, Vol. XVI (1955), 36-41.
- Copi, I. M. C. C. Elgot, and J. B. Wright, «Realization of Events by Logical Nets», *Journal for the Association of Computing Machinery*, Vol. V (1958), 181-196.
- Danto, Arthur C., «On Consciousness in Machines», in *Dimensions of Mind*, Sidney Hook, ed., pp. ISO-187. Nueva York: New York University Press, 1960.
- Davis, Martin, *Computability and Unsolvability*. Nueva York: McGraw-Hill Book Company, Inc., 1958.
- Friedberg, R. M., «A Learning Machine», Parte I, *IBM Journal of Research and Development*, Vol. II (1958) 2-13.
- Gelernter, H. L., «Realization of a Geometry-Proving Machine», *Proceedings of the International Conference on Information Processing*. Paris: UNESCO House, 1959.
- George, F. H., «Logic and Behavior», *Science News*, Vol. XLV (1957), 46-60.
- Godel, Kurt, «Uber Formal Unentscheidbare Satze der Principia Mathematica und verwandter Sys-teme», Parte I, *Monatshefte fur Mathematik und Physik*, Vol. XXXVIII (1931), 173-189.
- Haldane, J. B. S., «The Mechanical Chess-Player», *British Journal for the Philosophy of Science*, Vol. III (1952), 189-191.
- Hanson, Norwood Russell, *Patterns of Discovery*. Londres: Cambridge University Press, 1958.
- Hartree, D. ;R., *Calculating Instruments and Machines*. Urbana: University of Illinois Press, 1949.
- Jefferson, G., «The Mind of Mechanical Man», Lister Oration for 1949, *British Medical Journal*, Vol. I (1949), 1105-1121.
- Kleene, Stephen Cole, «General Recursive Functions of Natural Numbers», *American Journal of Mathematics*, Vol. LVII (1935), 153-173, 219-244.
- , *Introduction to Metamathematics*. Princeton: D. Van Nostrand Co., Inc., 1952.
- Lacey, A. R., «Men and Robots», *Philosophical Quarterly*, Vol. X (1960), 61-72.
- Lovelace, Mary Caroline, Countess of, *Scientific Memoirs*, R. Taylor, ed., Vol. III (1842), pp. 691-731.
- Lucas, J. R., «The Lesbian Rule», *Philosophy*, Vol. XXX (1955), 195-213.
- , «Minds, Machines, and Gödel», *Philosophy*, Vol. XXXVI (1961), 112-127.
- , «On Not Worshipping Facts», *Philosophical Quarterly*, Vol. VIII (1958), 144-156.
- MacKay, D. M., «The Epistemological Problem for Automata», *Automata Studies*, pp. 235-251. Princeton: Princeton University Press, 1956.
- , «Mentality in Machines», *Proceedings of the Aristotelian Society*, Supp. Vol. XXVI (1952), 61-86.

- , «Mind-Like Behavior in Artifacts», *British Journal for the Philosophy of Science*, Vol. II, No. 6 (1951), 110.
- , «Operational Aspects of Intellect», *Proceedings of the Symposium on Mechanization of Thought Processes*, D. V. Blake and A. M. Uttley, eds., London; Her Majesty's Stationery Office, 1959.
- McCulloch, W. S., «The Brain as a Computing Machine», *Electrical Engineering*, Vol. LXVIII (1949), 492-497.
- , and J. E. Pfeiffer, «Of Digital Computers Called Brains», *Scientific Monthly*, Vol. LXIX (1949), 368-376.
- Mays, W., «The Hypothesis of Cybernetics», *British Journal for the Philosophy of Science*, Vol. II, No. 7 (1951), 249-250.
- Minsky, Marvin L., «Heuristic Aspects of the Artificial Intelligence Problem», Group Report 34-55; Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, December 1956.
- , «A Selected Descriptor-Indexed Bibliography to the Literature on Artificial Intelligence», *IRE Transactions on Human Factors in Electronics* (1961), 39-55.
- , «Steps Toward Artificial Intelligence», *Proceedings of the IRE*, Vol. XLIX, No. 1 (1961), 8-30.
- Nagel, Ernest and J. R. Newman, *Godel's Proof*. Nueva York: New York University Press, 1958.
- Newell, A., J. C. Shaw, and H. A. Simon, «Elements of a Theory of Human Problem Solving», *Psychological Review*, Vol. LXV (1958), 151-166.
- Newell, A. and H. A. Simon, *The Simulation of Human Thought*, Report No. P-1734. Santa Monica, Cal.: The RAND Corporation, 1959.
- Oppenheim, Paul and Hilary Putnam, «Unity of Science as a Working Hypothesis», in *Concepts, Theories, and the Mind-Body Problem*, H. Feigl, G. Maxwell and M. Scriven, eds., Minnesota Studies in the Philosophy of Science, Vol. II. Minneapolis: University of Minnesota Press, 1958.
- Pitts, W. and W. S. McCulloch, «How We Know Universals», *Bulletin for Mathematical Biophysics*, Vol. IX (1947), 127-147.
- Popper, K. R., «Indeterminism in Quantum Physics and Classical Physics», *British Journal for the Philosophy of Science*, Vol. I (1951), 179-188.
- Putnam, Hilary, «The Analytic and the Synthetic», in *Scientific Explanation, Space, and Time*, H. Feigl and G. Maxwell, eds., Minnesota Studies in the Philosophy of Science, Vol. III. Minneapolis: University of Minnesota Press, 1962.
- Quine, Willard Van Orman, «The Scope and Language of Science», *British Journal for the Philosophy of Science*, Vol. VIII (1957), 1-17.
- Rashevsky, N., «Is the Concept of an Organism as a Machine a Useful One?», *Scientific Monthly*, Vol. LXXX (1955), 32-35.
- Richards, P. I., «On Game-Learning Machines», *Scientific Monthly*, Vol. LXXIV (1959), 201-205.

- Rogers, H., *Theory of Recursive Functions and Effective Computability* (mimeographed), Massachusetts Institute of Technology, 1957.
- Rorty, Amelie O., «Slaves and Machines», *Analysis*, Vol. XXII (1962), 118-120.
- Rosenbloom, Paul C., *Elements of Mathematical Logic*. Nueva York: Dover Publications, Inc., 1951.
- Ross, T., «The Synthesis of Intelligence-Its Implications», *Psychological Review*, Vol. XLV (1938), 185-189.
- Rosser, B., «Extensions of Some Theorems of Godel and Church», *Journal of Symbolic Logic*, Vol. I (1936), 87-91.
- Ryle, Gilbert, *The Concept of Mind*. Londres: Hutchinson & Co., Ltd., 1949.
- Samuel, A. L., «Some Studies in Machine Learning Using the Game of Checkers», *IBM Journal of Research and Development*, Vol. III (1959), 210-229.
- Scriven, Michael, «The Compleat Robot: A Prolegomena to Androidology», in *Dimensions of Mind*, Sidney Hook, ed., pp. 118-142. Nueva York: New York University Press, 1960.
- Selfridge, O. G., «Pandemonium: A Paradigm for Learning», Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, 1960.
- Shannon, C. E., «A Chess-Playing Machine», *Scientific American*, Vol. CLXXXII (1950), 48-51.
- and J. McCarthy, eds., *Automata Studies*. Princeton: Princeton University Press, 1956.
- Simmons, P. L. and R. F. Simmons, *The Simulation of Cognitive Behavior, II: An Annotated Bibliography*. Santa Monica, Cal.: System Development Corporation, SP-590/002/00, 1961.
- , «The Simulation of Cognitive Processes: An Annotated Bibliography», *IRE Transactions on Electronic Computers*, EC-10, No. 3 (1961), 462-483.
- Smart, J. J. C., «Godel's Theorem, Church's Theorem and Mechanism», *Synthese*, Vol. XIII, No. 2 (1961), 105-110.
- , «Incompatible Colors», *Philosophical Studies*, Vol. X (1959), 39-42.
- , «Sensations and Brain Processes», *The Philosophical Review*, Vol. LXVIII (1959), 141-156.
- Travis, L. ;E., «Observing how Humans Make Mistakes to Discover how to Get Computers to Do Likewise», Santa Monica, Cal.: System Development Corporation, SP-776, 1962.
- Turing, A. M., «On Computable Numbers, with and Application to the Entscheidungsproblem», *Proceedings of the London Mathematical Society*, Vol. XLII (1937), 230-265.
- von Neumann, John, *The Computer and the Brain*. New Haven: Yale University Press, 1958.
- , «The General and Logical Theory of Automata», *Cerebral Mechanisms in Behavior*, The Hixon Symposium, 1-31, L. A. Jefferies, ed. Nueva York: John

- Wiley & Sons, Inc., 1951.
- Wilkes, M. V., «Can Machines Think?» *Proceedings of the IRE*, Vol. XLI (1953), 1230-1234.
- Widsom, J. O., «The Hypothesis of Cybernetics», *British Journal for the Philosophy of Science*, Vol. II, No. 5 (1951-52), 1-24.
- , «A New Model for the Mind-Body Relationship», *British Journal for the Philosophy of Science*, Vol. II, No. 8 (1951-52), 295-301.
- , R. J. Spilsbury, and D. M. MacKay, «Symposium: Mentality in Machines», *Proceedings of the Aristotelian Society*, Supl. Vol. XXVI (1952), 1-61.
- Wright, M. A., «Can Machines Be Intelligent?» *Process Control and Automation*, Vol. VI (1959), 2-6.
- Ziff, Paul, *Semantic Analysis*. Ithaca: Cornell University Press, 1960.



ALAN ROSS ANDERSON (1925-1973), fue un lógico estadounidense y profesor de filosofía en las universidades de Yale y Pittsburgh en los Estados Unidos de América.

Obras: Some nasty problems in the formal logic of ethics. *Nous* I(4): 345-60, 1967; *Entailment: The Logic of Relevance and Necessity*. Vol. 1. Princeton: Princeton Univ. Press, 1975; *Entailment: The Logic of Relevance and Necessity*. Vol. 2. Princeton: Princeton Univ. Press, 1992.



ALAN TURING (23 de junio de 1912-7 de junio de 1954), fue un matemático, lógico, informático teórico, criptógrafo, filósofo, biólogo teórico, maratoniano y corredor de ultradistancia británico. Es considerado uno de los padres de la ciencia de la computación y precursor de la informática moderna.

Obras: Alan Turing, *On computable numbers, with an application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, Series 2, 42 (1936); *Intelligent Machinery* (1948); *Computing Machinery and Intelligence* (1950).



MICHAEL SCRIVEN (Gran Bretaña, 1928) es un académico estadounidense, titulado en matemática y doctorado en filosofía.

Obras: *The foundations of science and the concepts of psychology and psychoanalysis*. Minnesota Studies in the Philosophy of Science, 1956; *Concepts, theories, and the mind-body problem*. Minnesota Studies in the Philosophy of Science, Vol. 2, 1958; *Scientific explanation, space, and time*. Minnesota Studies in the Philosophy of Science, Vol. 3. , 1962; *Computers and comprehension*. RAND monograph, 1964; *Applied logic: An introduction to scientific method*. Behavioral Research Laboratories, 1965; *Primary philosophy*. McGraw-Hill, 1966; *Morality*. Social Science Education Consortium, 1966; *Perspectives of Curriculum Evaluation*. American Educational Research Association Monograph Series on Curriculum Evaluation, Vol. 1, 1967; *Evaluation Thesaurus*. Sage, 1991.



JOHN RANDOLPH LUCAS FBA (18 de junio de 1929 – 5 de abril de 2020) fue un filósofo británico.

Obras: *Principles of Politics*, 1966; *The Concept of Probability*, 1970; *The Freedom of the Will*, 1970; *The Nature of Mind* 1972; *A Treatise on Time and Space*, 1973; *Essays on Freedom and Grace*, 1976; *Space, Time and Causality*, 1985; *Conceptual Roots of Mathematics*, 1999.



KEITH R. GUNDERSON (1935 – octubre 14 de 2003) fue un filósofo estadounidense y profesor de la Universidad de Minnesota durante 40 años.

Obras: *Language, mind, and knowledge*, 1975; *Mentality and Machines*, 1985.



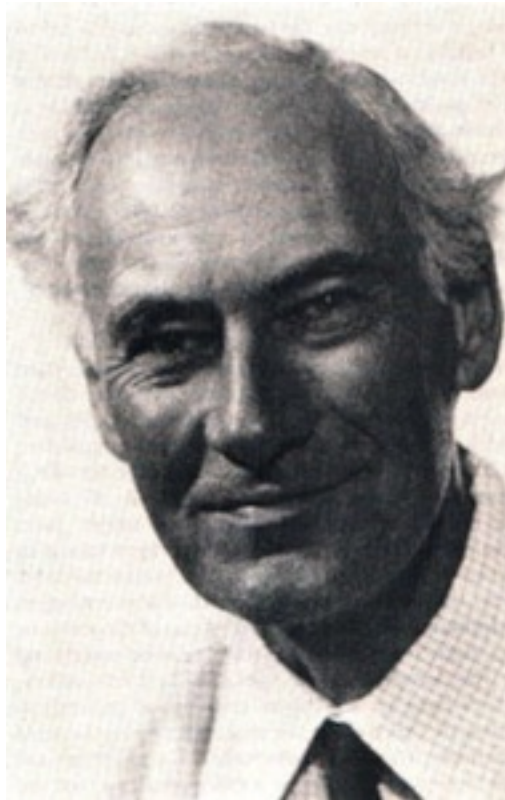
HILARY WHITEHALL PUTNAM (Chicago, Illinois, Estados Unidos, 31 de julio de 1926 - Arlington, Massachusetts, Estados Unidos, 13 de marzo de 2016) fue un filósofo, matemático e informático teórico estadounidense.

Obras: *Racionalidad y metafísica*, 1985; *Razón, verdad e historia*, 1988; *Representación y realidad: Un balance crítico de funcionalismo*, 1990; *Mente, lenguaje y realidad*, 2012; *Ética sin ontología*, 2013.



(ROBERT) PAUL ZIFF (22 de octubre de 1920, New York, Estados Unidos – 9 de enero de 2003, Chapel Hill, North Carolina) fue un artista y filósofo estadounidense especializado en semántica y estética.

Obras: *Art and the Object of Art*, 1951; *The Task of Defining a Work of Art*, 1953; *Reasons in Art Criticism*, 1958; *About Behaviourism*, 1958; *The Feelings of Robots*, 1959; *On What a Painting Represents*, 1960; *Semantic Analysis*, 1960; *On Understanding Understanding Utterances*, 1964; *The Simplicity of Other Minds*, 1965; *On H.P. Grice's Account of Meaning*, 1967.



JOHN JAMIESON CARSWELL “JACK” SMART **citado como** J. J. Smart (16 de septiembre de 1920 – 6 de octubre de 2012) fue un filósofo australiano de origen británico.

Obras: *An Outline of a System of Utilitarian Ethics*, 1961; *Philosophy and Scientific Realism*, 1963; *Problems of Space and Time*, 1964; *Between Science and Philosophy: An Introduction to the Philosophy of Science*, 1968; *Utilitarianism : For and Against*, 1973; *Ethics, Persuasion and Truth*, 1984; *Essays Metaphysical and Moral*, 1987; *Atheism and Theism*, 1996.



RODERICK NINIAN SMART (6 de mayo de 1927, Cambridge, Reino Unido – 29 de enero de 2001, Lancaster, Reino Unido) fue un filósofo británico.

Obras: *Reasons and Faiths: An Investigation of Religious Discourse, Christian and non-Christian*, 1958; *World Religions: A Dialogue*, 1960; *Secular Education and the Logic of Religion*, 1968; *Historical Selections In the Philosophy of Religion*, 1962; *Worldviews: Crosscultural Explorations of Human Belief*, 1981.

Notas

Maquinaria computadora e inteligencia

[*] «Maquinaria computadora e inteligencia», «Mind», Vol. LIX, n.º 236, 1950.
Publicado en este volumen con autorización de «Mind». (N. del E.) <<

[*] Se refiere a la primera computadora digital con programa almacenado instalada en la Universidad de Manchester en 1949. (N. del T.). <<

[1] Posiblemente es una consideración herética. Santo Tomás de Aquino, en *Summa Theologica*, citado por Bertrand Russell, *A History of Western Philosophy* (Simon and Schuster, Nueva York, 1945), pág 458, dice que Dios no puede hacer un hombre sin alma. Pero esto no sería una limitación real de su poder, sino tan sólo consecuencia del hecho de que el alma del hombre es inmortal y, por lo tanto, indestructible. <<

[*] Los nombres en cursiva de autores corresponden a trabajos citados en la bibliografía. (N. del E.) <<

[*] Ciudadana norteamericana, ciega y sorda de nacimiento, que logró una rehabilitación sorprendente y llegó a pronunciar conferencias. (N. del T.) <<

[2] O más bien «programado», ya que la máquina infantil se programaría en una computadora digital, aunque no tendría que aprender el sistema lógico. <<

[3] Compárese con la afirmación de lady Lovelace (pág. 35), en la que no consta la palabra «sólo». <<

El concepto mecánico de mente

[*] «El concepto mecánico de mente», «Mind», Vol. LXII. n.º 246, 1953. Reeditado con la autorización del autor y del editor de «Mind». (N. del E.) <<

[1] Léase: «Inconsciencia/Conciencia». <<

[*] En inglés ser y estar se expresan mediante una sola forma verbal. (N. del T.) <<

Mentes, máquinas y Gödel

[*] «Mentes, máquinas y Gödel», «Philosophy», Vol. XXXVI (1961). Reeditado con autorización del autor y del editor de «Philosophy». (N. del E.) <<

[1] Cf. A. M. Turing, «Maquinaria computadora e inteligencia» (primer cap. de este libro); K. R. Popper, «Indeterminism in Quantum Physics and Classical Physics»; Paul Rosenbloom, *Elements of Mathematical Logic*, págs. 207 y ss.; Ernest Nagel y James R. Newman, *Godel's Proof*, págs. 100 y ss.; y Hartley Rogers, *Theory of Recursive Functions and Effective Computability*, págs. 152 y ss. (todos ellos citados en la bibliografía al final de este libro). <<

[2] La prueba original de Gödel es aplicable; cf. los párrafos iniciales 1 y 6 de sus conferencias en el Instituto de Estudios Avanzados, Princeton, N. J., 1934. <<

[3] Para una argumentación en la misma línea, cf. los dos artículos de este mismo autor señalados en la bibliografía. <<

[4] En conversación privada. <<

[5] En *Theory of Recursive Functions*, op cit. <<

[6] La prueba original de Gödel es aplicable si la regla es susceptible de generar un tipo rudimentario, capaz de repetirse, de fórmulas adicionales; cf. párrafos iniciales 1 y 6 de sus conferencias en el Instituto de Estudios Avanzados. De hecho, basta con que el tipo sea capaz de repetirse indefinidamente. Cf. *Extensions of some Theorems of Gödel and Church*, de Barkley Rosser, en «Journal of Symbolic Logic», vol. I (1936), 87 y ss. <<

[7] *Ibid.*, pág. 154. <<

[8] En conversación privada. <<

El juego de la imitación

[*] «El juego de imitación», publicado en «Mind». Reproducido aquí con la autorización del autor y del editor de «Mind». (N. del E.) <<

[1] Cf. «El concepto mecánico de mente», Michael Scriven (segundo capítulo de este libro) y «The Compleat Robot: A Prolegomena to Andriology», en *Dimensions of Mind*, Sidney Hook, ed., New York University Press, Nueva York, 1960. Así como la observación que hace Paul Ziff en «El sentir de los robots» (sexto capítulo de este libro) y otros autores: por ejemplo, C. E. Shannon y J. McCarthy en el prólogo a *Autómata Studies*, Princeton University Press, Princeton, 1956. <<

Mentes y máquinas

[*] «Mentes y máquinas. Dimensions of Mind: A Symposium», Sidney Hook ed. New York University Press, Nueva York, 1960. Reproducido aquí con autorización del autor y de la New York University Press. (N. del E.) <<

[1] Por manifestación «desviada» entendemos la que se desvía de la regularidad semántica (en el lenguaje natural correspondiente). Hemos tomado el término de Paul Ziff, *Semantic Analysis*, Cornell University Press, Ithaca, N. Y., 1960. <<

[2] Cf. Rudolf Carnap, «The Interpretation of Physics», en *Readings in the Philosophy of Science*, H. Feigl y M. Brodbeck, eds., Appleton-Century-Crofts, Inc., Nueva York, 1953; y «The Methodological Character of Theoretical Concepts», en *Foundations of Science and the Concepts of Psychology and Psychoanalysis*, H. Feigl y M. Scriven eds., *Minnesota Studies in the Philosophy of Science*, Vol. I, University of Minnesota Press, Minneapolis, 1956. Este modelo de teoría científica está, en mi opinión, excesivamente simplificado para que resulte de utilidad general; no obstante, la excesiva simplificación no afecta a nuestra argumentación. <<

[3] La cuestión quedó planteada en el trabajo de Willard Van Orman Quine, «The Scope and Language of Science», «British Journal for the Philosophy of Science», Vol. VIII, 1957. <<

[4] Cf. Paul Ziff, «El sentir de los robots»; y J. J. C. Smart, «El profesor Ziff y los robots» (más adelante, págs. 151 y ss.). Ziff me ha comunicado que no entendía por «robot» una «máquina que aprende», como el propuesto por Smart, y que, desde luego, las consideraciones expuestas en su trabajo no se referían necesariamente a este tipo de máquina (si, de hecho, puede calificarse de «máquina»). Sobre si la cuestión de si «esta máquina piensa (siente, etc.)» es desviada o no, hay que tener en cuenta tanto la cuestión planteada por Ziff (que lo importante no es si la afirmación es desviada o no, sino si es o no desviada por motivos no triviales), así como la distinción «diacrónica-sincrónica» de que se habla en el apartado 5. <<

[5] Me inclino en particular por la postura adoptada por Smart en «Incompatible Colors», *Philosophical Studies*, Vol. X, 39 y ss., 1959 y «Sensations and Brain Processes», «*Philosophical Review*», Vol. LXVni, 141 y ss., 1959. No obstante, véase las consideraciones lingüísticas en el apartado 5. <<

[6] Para más detalles, cf. Martin Davis, *Computability and Unsolvability*, McGraw-Hill Book Company, Inc., Nueva York, 1958, y Stephen Cole Klecne, *Introduction to Mathematics*, D. Van Nostrand Co., Inc., Princeton, 1952. <<

[7] Adoptamos esta terminología de Kleene, *op. cit.*, para diferenciarla de la empleada por Davis y Turing. <<

[8] Esta generalización se postula en Davis, *op. cit.*, en donde se emplea para definir la relativa repetición. <<

[9] Esta afirmación es una variante de la tesis de Church (según la cual la repetición equivale a una computabilidad efectiva). <<

[10] Aquí hay que suponer que toda la frase «Estoy en el estado A» equivale a un solo símbolo del alfabeto de la máquina. <<

[*] En inglés existe una sola forma verbal (*to know*) para saber y conocer. (N. del T.)
<<

[¹¹] Por ejemplo, sé que el sol se halla a unos 149 millones de kilómetros de la Tierra, pero no dispongo de la *evidencia* de que así es. En realidad, ni siquiera recuerdo dónde lo aprendí. <<

[12] En realidad, resultaría imposible determinar si «Jones sabe que le duele» es desviada o no sin reformular primero la condición evidencial para evitar la objeción de la nota 11 (si es que puede reformularse para conservar en algo esta condición). Sin embargo, creo que la discusión que antecede indica por qué se puede *desear* que esta frase sea desviada. <<

[13] En realidad, no era necesario añadirles ningún «órgano sensorial»; las computadoras se autocomprobaban mediante «experimentos cruciales consigo mismas» (realizando, por ejemplo, determinados tests de computación y comparando los resultados con los resultados correctos que les habían introducido). <<

[14] En el sentido de Paúl Oppenheim e Hilary Putnam, «Unity of Science as a Working Hypothesis», en *Concepts, Theories, and the Mind-Body Problem*, H. Feigl, G. Maxwell y M. Scriven, eds., *Minnesota Studies in the Philosophy of Science*, Vol. II, University of Minnesota Press, Minneapolis, 1958; no en el sentido «epistemológico» asociado a los escritos de Carnap sobre «fisicalismo». <<

[15] Hilary Putnam, «The Analytic and the Synthetic», en *Scientific Explanation, Space and Time*, H. Feigl y G. Maxwell eds., *Minnesota Studies in the Philosophy of Science*, Vol, III, University of Minnesota Press, Minneapolis, 1962. <<

[16] Norwood Russell Hanson, *Patterns of Discovery*, Cambridge University Press, Londres, 1958. <<

[17] Cf. *Ibid.* <<

[18] Cf. La discusión sobre geometría en Putnam, *op. cit.* <<

[19] La lingüística diacrónica estudia el lenguaje a través de su evolución en el tiempo; la lingüística sincrónica trata sólo de describir el lenguaje en un momento determinado. <<

[20] Para una discusión pormenorizada, cf. Ziff, *Semantic Analysis*, *op. cit.* Quedo profundamente agradecido a Ziff por procurarme este libro y por sus informaciones privadas sobre el tema. El apartado 5 de este trabajo representa, en parte, la influencia de Ziff (particularmente en lo que atañe al uso de la distinción entre sincrónico y diacrónico) y, en parte, el desarrollo de ciertas ideas de Putnam, *op. cit.*, sobre la materia. <<

[21] Término tomado de *Semantic Analysis* de Ziff. <<

[22] La *distribución* de una palabra equivale a la serie de frases en que se produce. <<

[23] En este caso «Estado mental es idéntico a estado cerebral» se emplea como equivalente a «El dolor es idéntico a la estimulación de fibras C». <<

[24] Argumentado en Putnam, *op. cit.* <<

[25] Por frase de este *tipo* no quiero decir literalmente *ejemplos de sustitución* de «el estado mental ψ es idéntico al estado cerebral ϕ ». Cf. nota 23. <<

[26] He tenido mis dudas al atribuir esta actitud a Black en base a sus observaciones durante el seminario. Pero, naturalmente, comprendo que no se le puede en rigor hacer responsable de observaciones efectuadas sobre la marcha. <<

[27] Empleo este término en el sentido de 3), no en el tradicional. <<

El sentir de los robots

[*] «El sentir de los robots», «Analysis», Vol. XIX, núm. 3, 1959. Reproducido aquí con la autorización del editor de «Analysis». (N. del E.) <<

[1] Cf. D. M. MacKay, «The Epistemological Problem for Automata», en *Automata Studies*, págs. 235 y ss., Princeton University Press, 1956. <<

[2] Cf. M. Scriven «El concepto mecánico de mente». (Reproducido en este volumen.)
<<

[3] De aquí en adelante prescindiré de esta posibilidad. <<

[4] Cf. «Maquinaria computadora e inteligencia». (Reproducido en este volumen). <<

[5] Cf. «Mentality in Machines», *Proceedings of the Aristotelian Society*, 61 y ss.,
supl. Vol. XXVI. 1952 <<

El profesor Ziff y los robots

[*] «El profesor Ziff y los robots», «Analysis», vol. XIX (1959). Reproducido aquí con la autorización del autor y del editor de «Analysis». (N. del E.) <<

Robots, S. A.

[*] «Robots, S. A.», «Analysis», vol. XIX, núm. 5, 1959. (Reproducido aquí con la autorización del autor y del editor de «Analysis»). <<

[1] En lo que sigue tanto Ziff como yo empleamos el vocablo «programa» en sentido bastante amplio. <<

[2] Pág. 151 de este libro. <<

[3] Pág. 156 de este libro. <<